CAN HYBRID-SPACE REPRESENTATION LEARNING ADVANCE MULTIMODAL FUSION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fusion—joint learning of representations from heterogeneous data—has shown significant potential in healthcare by integrating imaging, text, and omics for comprehensive analysis. However, prior methods face two major limitations: (1) inadequate modeling of complex hierarchical structural dependencies at high computational cost, which *constrains the performance–efficiency trade-off*; and (2) *limited robustness to missing modalities*. These limitations restrict their applicability in resource-constrained medical `AI`. To address these challenges, we introduce **Hybrid Space Attention Refiner** (`HySAR`), a plug-and-play cross-modal refiner that jointly exploits Euclidean and non-Euclidean manifolds to capture complex hierarchical structural dependencies for learning robust shared representations. This, in turn, improves multimodal fusion performance and missing-modality robustness while `HySAR`'s channel-wise geometric refinement and complementary multi-scale information learning further reduce computational cost. Across 10 multimodal benchmarks, `HySAR`-integrated learners achieve $\approx$ 4.8-point on average improvements over baselines, up to $\approx$ 84.1% parameter and $\approx$ 88.0% `FLOP` reductions, and up to $\approx$ 8.7-point gains in missing-modality robustness, enabling efficient and robust clinical prediction. Code — HySAR

1 INTRODUCTION

Multimodal fusion is a powerful framework for diverse applications—including disease analysis and patient survival prediction—by learning holistic shared representations from heterogeneous data sources, e.g., imaging, multi-omics, and clinical data (Dhar et al., 2026b). Such representation learning can highlight salient regions and modality-specific patterns across diverse modalities, thereby strengthening multimodal fusion effectiveness (Dhar et al., 2025b; Hemker et al., 2025; Dhar et al., 2026b). Despite rapid progress, existing fusion architectures still face two key challenges—**performance–efficiency trade-off** (Fig. 1) and **robustness to missing modalities** (Fig. 5)—which are especially crucial for resource-constrained medical `AI` applications.

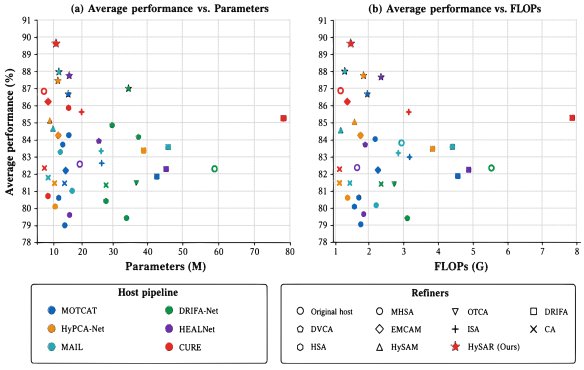

Figure 1: Performance–efficiency trade-off of cross-modal refiners in `ViT-Tiny`-based multimodal fusion pipelines across six medical benchmarks (see Fig. 4 and Fig. 12 (App. F.1)). Overall performance is the mean of `C-index`, `ACC`, and `AUC`. Colors, marker shapes, and stars denote host pipelines, refiners, and `HySAR`, respectively.

First, prior multimodal fusion pipelines (Dhar et al., 2025a; Hemker et al., 2024; Xu & Chen, 2023; Dhar et al., 2026a) often adopt Euclidean-space-aware attention mechanisms organized as cascaded modules (Dhar et al., 2026b). Although effective for spatial and channel refinement, these cascaded designs can progressively attenuate discriminative evidence across inter-module transitions, remain limited in modeling coarse-to-fine hierarchical structural dependencies, and incur substantial computational overhead (Dhar et al., 2026b; Hemker et al., 2024; Dhar et al., 2026c). Conversely, non-Euclidean-only refiners (Dhar et al., 2026b) capture hierarchy-aware cues but remain limited in preserving fine-grained Euclidean evidence. Exploring either geometry in isolation therefore creates a representational–computational bottleneck: limited representation quality constrains downstream performance (Figs. 1–2) (Hemker

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

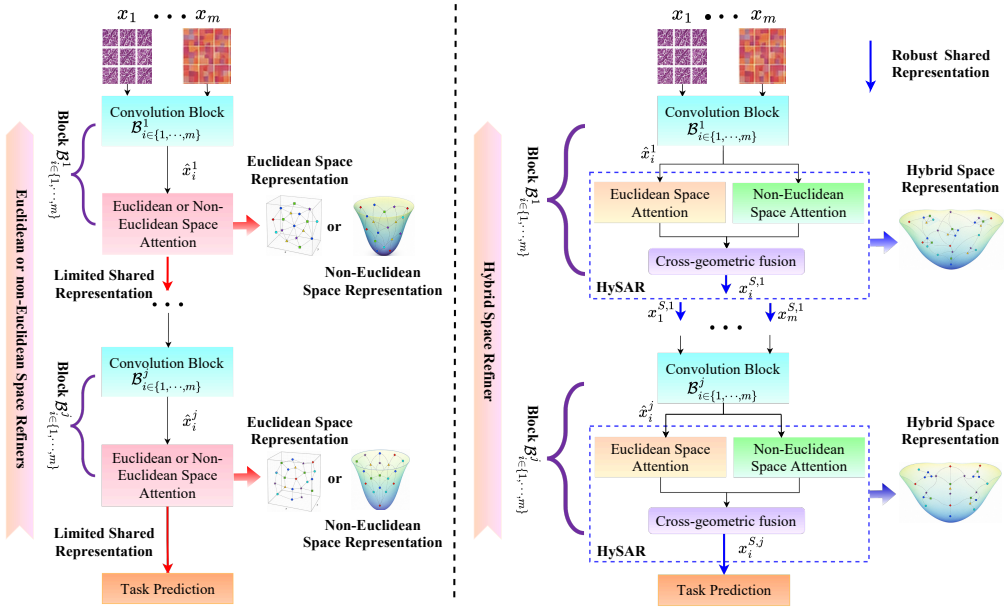

Figure 2: **Single-geometry vs. hybrid-space cross-modal refiners for shared representation learning in multimodal fusion. Left:** Euclidean-space cross-modal refiners incorporated into multimodal fusion learners—e.g., EMCAM within MAIL (Dhar et al., 2025a)—capture fine-grained dependencies, whereas the non-Euclidean-space refiner HySAM within CURE (Dhar et al., 2026b) captures hierarchy-aware structural dependencies. These designs remain confined to a single geometry and leave the complementary strengths of both spaces underexplored, thereby limiting shared-representation expressiveness. **Right:** Plug-and-play HySAR cross-fuses complementary Euclidean and non-Euclidean evidence across modalities to capture complex coarse-to-fine hierarchical structural cues, thereby learning robust shared representations $x_i^{S,j}$ that strengthen multimodal fusion effectiveness (see Table 1 and Fig. 3).

et al., 2024), while cost-intensive attention architectures (Dhar et al., 2025b; Joze et al., 2020) increase computational burden and restrict applicability in resource-constrained multimodal AI . These limitations constrain simultaneous improvements in performance and efficiency (Dhar et al., 2026b) (**Challenge #1**). A second challenge arises from the common assumption in prior multimodal fusion methods that all modalities are available during both training and inference, making them vulnerable to missing-modality scenarios (Swamy et al., 2023) (**Challenge #2**).

To address these challenges, we introduce **Hybrid Space Attention Refiner** (HySAR)—a plug-and-play cross-modal refiner that replaces Euclidean- or non-Euclidean-space refiners in multimodal fusion (Fig. 2), thereby offering a robust and efficient solution for resource-constrained medical AI. As shown in Fig. 3, the core components of HySAR are **Multi-scale Channel Information Learning Attention** (MILA), which preserves fine-grained multi-scale spatial–channel contexts, and **Dual Hyperbolic Fusion Attention** (DUHA), which learns complementary hierarchy-aware responses. A cross-geometric fusion subsequently integrates both geometric responses across modalities to capture diverse complex coarse-to-fine hierarchical structural dependencies. This, in turn, facilitates learning robust shared representations that improve multimodal fusion performance and missing-modality robustness. Moreover, HySAR's channel-wise geometric refinement and multi-scale information learning further reduce computation, allowing such complementary designs to improve the performance–efficiency trade-off (Fig. 1 and Tables 1–2). The main contributions of our work are:

1. We introduce HySAR, a plug-and-play cross-modal refiner that addresses challenges—the *performance–efficiency trade-off* and *robustness to missing modalities*—making it well suited for resource-constrained medical AI.
2. We propose MILA (Euclidean-space attention) for learning multi-scale spatial–channel cues and DUHA (non-Euclidean-space attention) for capturing complementary hierarchy-aware dependencies, showing their cross-geometric integration yields robust shared representations.
3. We conduct extensive experiments on ten medical and natural-vision multimodal benchmarks, demonstrating consistent improvements in *performance* and *missing-modality robustness* over state-of-the-art methods across diverse multimodal fusion learners.

Related work. We include an extensive discussion of related work in Appendix A.

2 METHOD

Problem formulation and method overview. Let $X = \{x_i\}_{i=1}^m$ denote a multimodal input comprising m heterogeneous modalities, where $x_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ denotes an image-like tensor and $x_i \in \mathbb{R}^{D_i}$ denotes a vector; y is the ground-truth label. We consider multimodal fusion pipelines (Dhar et al., 2025b; 2026c) composed of J blocks, where $j \in [J] = \{1, \ldots, J\}$ and J is the total number of blocks. Each block contains two functional components: a modality-specific convolution block $\mathcal{B}_i^j$, e.g., `ERLA` (Dhar et al., 2026c), and a cross-modal refiner $\mathcal{R}_\phi^j$, e.g., `DRIFA` (Dhar et al., 2025b) or `EMCAM` (Dhar et al., 2025a). For modality i at block j, $\mathcal{B}_i^j$ encodes the input representation x_i^j into the contextual feature $\hat{x}_i^j = \mathcal{B}_i^j(x_i^j; \eta_i^j)$, where η_i^j denotes its parameters. These features are subsequently processed by the cross-modal refiner to learn shared representations:

$$\{x_i^{S,j}\}_{i=1}^m = \mathcal{R}_\phi^j\left(\{\hat{x}_i^j\}_{i=1}^m\right). \tag{1}$$

The shared representations of the final block are used for prediction: $\hat{y} = \mathcal{F}(\{x_i^{S,J}\}_{i=1}^m; \beta)$, where $\mathcal{F}(\cdot; \beta)$ denotes the multimodal prediction function parameterized by β. Our objective is to learn robust shared representations that improve performance while remaining computationally efficient and reliable under missing modalities. Since the cross-modal refiner is central to learning such representations, we focus on redesigning this component. We therefore minimize such fusion objective

$$\mathcal{L}_{\texttt{MFL}}(X, y; \Theta, \Lambda) = \sum_{i=1}^m \lambda_i \, \mathcal{L}_i\left(\mathcal{F}(x_i^{S,J}; \beta), y\right), \tag{2}$$

where $\Theta = \{\eta, \phi, \beta\}$ denotes the trainable parameters and $\Lambda = \{\lambda_i\}_{i=1}^m$ controls the contribution of each modality-specific loss $\mathcal{L}_i$. To satisfy these objectives, existing multimodal fusion architectures, such as `DRIFA-Net` (Dhar et al., 2025b) and `MAIL` (Dhar et al., 2026c), commonly instantiate $\mathcal{R}_\phi^j$ in either Euclidean geometry or non-Euclidean manifolds. Euclidean refiners (e.g., `EMCAM` (Dhar et al., 2025a)) effectively capture fine-grained dependencies but are less suited to modeling complex hierarchical structural cues (Peng et al., 2021). Conversely, non-Euclidean refiners (e.g., `HySAM` (Dhar et al., 2026b)) emphasize hierarchical structural cues but do not explicitly preserve fine-grained spatial evidence. Prior refiners therefore often explore these geometries in isolation, leaving their complementary strengths underexplored and restricting shared-representation expressiveness. This, in turn, constrains multimodal fusion effectiveness (see Table 1).

We bridge this gap by introducing the **Hybrid Space Attention Refiner** (`HySAR`), a plug-and-play replacement for the cross-modal refiner that learns robust shared representations. It jointly exploits Euclidean and non-Euclidean evidence to preserve multi-scale spatial–channel dependencies while capturing complementary coarse-to-fine hierarchical structural cues. `HySAR` is implemented through three coordinated modules: **(i) Multi-scale Channel Information Learning Attention** (`MILA`) for learning multi-scale information; **(ii) Dual Hyperbolic Fusion Attention** (`DUHA`) for capturing diversified hierarchy-aware responses; and **(iii) Hybrid-Space Multi-Cross Fusion** (`HFuse`) for integrating both geometric properties across modalities. This design improves the performance–efficiency trade-off (Tables 1–2; Fig. 1) and missing-modality robustness (Fig. 5).

An overview of `HySAR` is shown in Fig. 3, and its pseudocode is provided in Algorithms 1–4 (see App. C). For each modality i, `HySAR` learns a hybrid-space attention map $A_i^{(H,j)}$ and uses it to obtain the robust shared representation $x_i^{S,j}$ defined in Eq. 1:

$$x_i^{S,j} = \hat{x}_i^j \odot A_i^{(H,j)} \odot \alpha_i^j, \text{ where } A_i^{(H,j)} = \texttt{HFuse}_i^j\left(A_i^{(D,j)}, \left\{A_r^{(E,j)}\right\}_{\substack{r=1 \\ r \neq i}}^m\right). \tag{3}$$

Here, $A_i^{(E,j)} = \texttt{MILA}(\hat{x}_i^j)$ and $A_i^{(D,j)} = \texttt{DUHA}(\hat{x}_i^j)$ denote the Euclidean-space and non-Euclidean-manifold attention maps for each modality, respectively; $r \neq i$ indexes the complementary modalities and α_i^j is a learnable parameter. The shared representation $x_i^{S,j}$ is forwarded to the subsequent block when $j < J$ and to the task-specific prediction block when $j = J$. The formulations of `MILA`, `DUHA`, and `HFuse` are described in the following subsections.

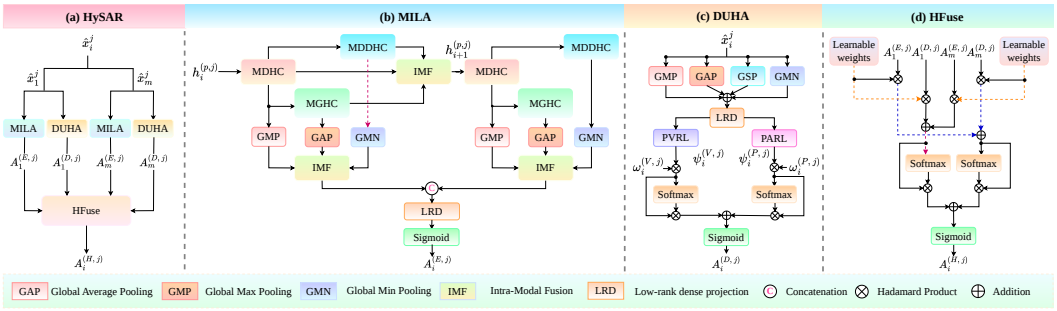

Figure 3: **Overview of the proposed HySAR. (a)** For modality i at block j, HySAR coordinates three core modules—MILA, DUHA, and HFuse—to learn the hybrid-space attention map $A_i^{(H,j)}$ for robust shared representation learning. **(b)** MILA applies p-th stage (i.e., $p \in \{1, 2\}$) multi-scale homogeneous refinement and channel-context aggregation to learn the Euclidean-space attention map $A_i^{(E,j)}$. **(c)** DUHA integrates the parallel PARL and PVRL responses through geometry-wise gating to learn the dual-hyperbolic attention map $A_i^{(D,j)}$, which captures complementary hierarchy-aware structural dependencies. **(d)** HFuse adaptively cross-fuses non-Euclidean evidence with Euclidean evidence from the complementary modalities to learn $A_i^{(H,j)}$, which recalibrates $\hat{x}_i^j$ into the robust shared representation $x_i^{S,j}$, thereby strengthens multimodal fusion effectiveness.

2.1 Multi-scale Channel Information Learning Attention

Motivation and overview. Euclidean-space attention learners, such as DRIFA (Dhar et al., 2025b) and EMCAM (Dhar et al., 2025a), capture fine-grained evidence for multimodal fusion but incur non-trivial computational cost, which limits their use in resource-constrained medical AI. We therefore design **Multi-scale Channel Information Learning Attention** (MILA), an efficient Euclidean-space representation learner that captures complementary coarse-to-fine channel responses from multi-scale spatial evidence, thereby strengthening representation learning within HySAR. As shown in Fig. 3(b), MILA is implemented through two phases: **(i) multi-scale homogeneous convolution streams** for fine-, mid-, and coarse-grained spatial information refinement, and **(ii) multi-scale channel context learning** for summarizing these responses into Euclidean-space channel attention maps. This design improves the performance–efficiency trade-off of HySAR-integrated multimodal fusion while preserving representational expressiveness.

As shown in Fig. 3(b), given input $h_i^{(1,j)} = \hat{x}_i^j$, MILA applies a p-th-stage multi-scale learning process, where $p \in \{1, 2\}$. Each stage refines its input representation and extracts its channel context, while the subsequent stage receives this refined output as its input representation. These contexts are aggregated to learn the Euclidean-space attention map. We discuss both phases below.

1. **Multi-scale homogeneous convolution streams.** We design Multi-scale homogeneous convolution streams by implementing three homogeneous streams: **(a) Multi-scale Depth-wise Homogeneous Convolutions** (MDHC), **(b) Multi-scale Group-wise Homogeneous Convolutions** (MGHC), and **(c) Multi-scale Dilated Depth-wise Homogeneous Convolutions** (MDDHC). These streams provide complementary multi-scale spatial inductive biases. Specifically, MDHC preserves local-to-coarse spatial evidence, MGHC captures structured spatial responses with intra-group channel interactions, and MDDHC enlarges the effective receptive field for broader contextual aggregation. They capture complementary local, group-structured, and context-expanded spatial evidence within a unified lightweight computation pattern.

   The MDHC, MGHC, and MDDHC streams apply depth-wise convolution (DWC), group-wise convolution (GWC), and dilated depth-wise convolution (DDC) bases, respectively. Each basis is evaluated using diverse kernel sizes $\mathcal{K} = \{1\times1, 3\times3, 5\times5\}$ to capture local, mid-range, and coarse spatial patterns. The resulting multi-scale responses are concatenated and channel-shuffled to promote inter-channel communication. A group point-wise convolution (GPC) subsequently consolidates these responses, while a residual connection preserves modality-specific evidence and stabilizes optimization. Since the three streams follow the same multi-scale computation template, we define the p-th-stage homogeneous basis for $\Omega \in \{\mathtt{DWC}, \mathtt{GWC}, \mathtt{DDC}\}$ as

$$\mathcal{B}_{i,\Omega}^{(p,j)}(z) = z + \mathtt{GPC}_{i,\Omega}^{(p,j)}\left(\mathtt{CS}_i^j\left(\theta_i^j\left[\left\{\Omega_{i,k}^{(p,j)}(z)\right\}_{k\in\mathcal{K}}\right]\right)\right), \tag{4}$$

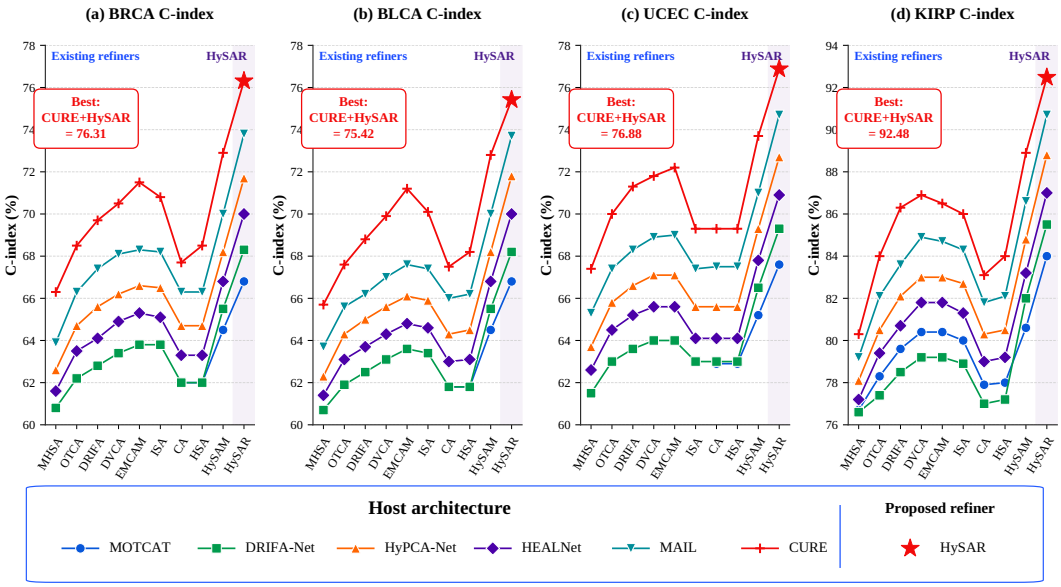

Figure 4: Performance comparison (`C-index`) of plug-and-play refiners integrated into `ViT-Tiny`-based multimodal fusion pipelines on the `BRCA`, `BLCA`, `UCEC`, and `KIRP` benchmarks. Each curve traces a host pipeline across replacement refiners; stars denote `HySAR`.

where $\theta_i^j[\cdot]$ denotes channel-wise concatenation and `CS` denotes channel shuffle. Accordingly, $\psi_i^{(p,j)} = \mathcal{B}_{i,\texttt{DWC}}^{(p,j)}$, $\upsilon_i^{(p,j)} = \mathcal{B}_{i,\texttt{GWC}}^{(p,j)}$, and $\tau_i^{(p,j)} = \mathcal{B}_{i,\texttt{DDC}}^{(p,j)}$ denote `MDHC`, `MGHC`, and `MDDHC`, respectively. For each stage p, `MDHC` first extracts the base multi-scale response $\varsigma_{*,i}^{(p,j)} = \psi_i^{(p,j)}(h_i^{(p,j)})$. `MGHC` and `MDDHC` then process this response in parallel to obtain $\lambda_i^{(p,j)} = \upsilon_i^{(p,j)}(\varsigma_{*,i}^{(p,j)})$ and $\xi_i^{(p,j)} = \tau_i^{(p,j)}(\varsigma_{*,i}^{(p,j)})$, respectively. Their responses are integrated through intra-modal fusion (`IMF`), defined as $\mathcal{I}(a,b,c) = \big(a+b+c\big) + \big(a-b-c\big)$, where the bracketed terms represent parallel additive and subtractive response paths. These paths are fused to learn the refined multi-scale representation and forwarded to the next `MILA` stage:

$$h_i^{(p+1,j)} = \varphi_i^{(p,j)} = \mathcal{I}\Big(\varsigma_{*,i}^{(p,j)}, \lambda_i^{(p,j)}, \xi_i^{(p,j)}\Big), \qquad p \in \{1,2\}. \tag{5}$$

2. **Multi-scale channel context learning.** We design Multi-scale channel context learning to summarize the stage-wise multi-scale responses using global max pooling (`GMP`), global average pooling (`GAP`), and global minimum pooling (`GMN`). For each stage p, these statistics capture salient responses, average contextual evidence, and complementary low-response structures, respectively; the resulting channel summaries are then refined using the `IMF` mechanism $\mathcal{I}(\cdot)$ to obtain the stage-wise diverse channel context.

$$\zeta_i^{(p,j)} = \mathcal{I}\Big(\texttt{GMP}\Big(\varsigma_{*,i}^{(p,j)}\Big), \texttt{GAP}\Big(\lambda_i^{(p,j)}\Big), \texttt{GMN}\Big(\xi_i^{(p,j)}\Big)\Big). \tag{6}$$

The resulting contexts are aggregated across the p-th stages to learn discriminative multi-scale channel dependencies. We then apply a low-rank dense projection to reduce computational cost, followed by a sigmoid $\sigma(\cdot)$ that learns the Euclidean-space channel attention map:

$$A_i^{(E,j)} = \sigma\Big(f_c^j\Big(\theta\zeta_i^{(p,j)}\Big)\Big) \qquad \text{s.t.} \qquad p \in \{1,2\}. \tag{7}$$

Here, θ denotes concatenation, $f_c^j(\cdot)$ denotes the low-rank dense projection. The learned $A_i^{(E,j)}$ provides multi-scale channel evidence for subsequent hybrid-space fusion within `HySAR`.

2.2 Dual Hyperbolic Fusion Attention

Motivation. Non-Euclidean manifold representation learning, e.g., through hyperbolic attention mechanisms (Ganea et al., 2018; Gulcehre et al., 2018; Peng et al., 2021), is well suited to modeling hierarchical structural dependencies. However, relying on an individual hyperbolic formulation may limit the capture of complex and diversified hierarchy-aware structural cues. This limitation is

particularly consequential in multimodal fusion, where complementary evidence spans different semantic levels and requires expressive geometry-aware calibration. Distinct hyperbolic formulations can induce complementary non-Euclidean responses, yet their unified representational strengths remain underexplored for multimodal fusion.

We bridge this gap by proposing **Dual Hyperbolic Fusion Attention** (DUHA), a core non-Euclidean geometry-aware module of HySAR that jointly exploits distinct hyperbolic responses within a unified attention formulation to capture diversified coarse-to-fine hierarchical structural cues. This strengthens non-Euclidean representation learning, which improves multimodal fusion effectiveness.

As shown in Fig. 3(c), given the modality-specific contextual feature $\hat{x}_i^j \in \mathbb{R}^{H_j \times W_j \times C_j}$, DUHA learns the dual-hyperbolic attention map $A_i^{(D,j)}$ through implementing two parallel streams: **(i) Poincaré-Aware Representation Learning** (PARL) and **(ii) Proper-Velocity Manifold Representation Learning** (PVRL). Before manifold-specific processing, $\hat{x}_i^j$ is summarized using diverse global channel statistics: global average pooling (GAP), global max pooling (GMP), global sum pooling (GSP), and global minimum pooling (GMN). Independent low-rank projections (Cheng et al., 2018) then reduce computational cost and learn formulation-specific channel descriptors:

$$\lambda_i^{(g,j)} = f_g^j\Big(\texttt{GAP}(\hat{x}_i^j) + \texttt{GMP}(\hat{x}_i^j) + \texttt{GSP}(\hat{x}_i^j) + \texttt{GMN}(\hat{x}_i^j)\Big), \qquad g \in \{P, V\}. \tag{8}$$

Here, P and V index the PARL and PVRL streams, respectively, while $f_P^j(\cdot)$ and $f_V^j(\cdot)$ denote their independently learned low-rank projections. Let $\psi_i^{(P,j)} = \texttt{PARL}(\lambda_i^{(P,j)})$ and $\psi_i^{(V,j)} = \texttt{PVRL}(\lambda_i^{(V,j)})$ denote the corresponding hyperbolic responses. Their interaction is formalized as

$$A_i^{(D,j)} = \sigma\left(\sum_{g\in\{P,V\}} \mathrm{softmax}_g\left(\omega_i^{(g,j)} \odot \psi_i^{(g,j)}\right) \odot \left(\omega_i^{(g,j)} \odot \psi_i^{(g,j)}\right)\right). \tag{9}$$

Here, $\omega_i^{(g,j)}$ denotes learnable channel-wise modulation parameters. We use softmax to gate the more informative hyperbolic formulation and sigmoid to learn the dual-hyperbolic attention map. We describe PARL and PVRL below.

1. **Poincaré-Aware Representation Learning.** We design Poincaré-Aware Representation Learning (PARL) to learn hierarchy-aware channel responses on the Poincaré ball (Nickel & Kiela, 2017; Ganea et al., 2018). As shown in Fig. 6 in App. B, given $\lambda_i^{(P,j)}$, PARL maps the Euclidean channel descriptor onto the Poincaré ball through the Poincaré exponential map (PoEM) at the origin with learnable curvature magnitude $c_P^j > 0$. The hyperbolic embedding modulates the original descriptor to produce $\psi_i^{(P,j)}$, which captures curvature-aware radial hierarchy while preserving channel-specific response magnitudes:

$$\psi_i^{(P,j)} = \texttt{PARL}\left(\lambda_i^{(P,j)}\right) = \tanh\left(\sqrt{c_P^j}\left\|\lambda_i^{(P,j)}\right\|_2\right)\frac{\lambda_i^{(P,j)}}{\sqrt{c_P^j}\left\|\lambda_i^{(P,j)}\right\|_2 + \varepsilon} \odot \lambda_i^{(P,j)}. \tag{10}$$

   Here, $\varepsilon = 10^{-6}$ ensures numerical stability, while $c_P^j = \mathrm{softplus}(k_P^j) + 10^{-3}$ maintains a positive curvature magnitude during optimization.

2. **Proper-Velocity Manifold Representation Learning.** We design Proper-Velocity Manifold Representation Learning (PVRL) by extending the Proper-Velocity (PV) manifold (Chen et al., 2026), rooted in Einstein's special relativity (Ungar, 2008), into an attention formulation that exploits the unconstrained coordinate representation of hyperbolic space to strengthen multimodal representation learning. As shown in Fig. 7 in App. B, given $\lambda_i^{(V,j)}$, PVRL applies the PV exponential map at the origin, an intrinsic PV fully connected transformation, and the corresponding logarithmic map. Accordingly, $\psi_i^{(V,j)}$ captures hierarchy-aware channel transformations within an unconstrained hyperbolic coordinate domain:

$$\psi_i^{(V,j)} = \texttt{PVRL}\left(\lambda_i^{(V,j)}\right) = \mathrm{Log}_0^{\mathrm{PV},K_V^j}\left(\mathrm{PVFC}_{K_V^j}^{C_j\to C_j}\left(\mathrm{Exp}_0^{\mathrm{PV},K_V^j}\left(\lambda_i^{(V,j)}\right)\right)\right). \tag{11}$$

   Here, $K_V^j = -(\mathrm{softplus}(k_V^j) + 10^{-3}) < 0$ denotes the learnable negative curvature, $\mathrm{Exp}_0^{\mathrm{PV},K_V^j}(\cdot)$ and $\mathrm{Log}_0^{\mathrm{PV},K_V^j}(\cdot)$ denote the closed-form PV exponential and logarithmic maps, respectively, and $\mathrm{PVFC}_{K_V^j}^{C_j\to C_j}(\cdot)$ denotes the intrinsic PV fully connected transformation.

Table 1: **Performance–efficiency trade-offs of single-geometry versus hybrid-space plug-and-play cross-modal refiners.** We evaluate each refiner when integrated into ResNet18-based multimodal fusion pipelines—HEALNet, MAIL, and CURE—across six medical benchmarks: BRCA, BLCA, UCEC, KIRP, HAM10000, and SIPaKMeD. Each host's first row uses its native refiner; each "+" row replaces only this refiner with a Euclidean-space, non-Euclidean-manifold, or hybrid-space HySAR module.

Datasets →	BRCA	BLCA	UCEC	KIRP	HAM10000		SIPaKMeD		#P	#F
HEALNet	64.52±5.99	66.47±4.00	65.14±3.02	81.90±5.62	95.58±1.43	95.27±1.91	93.91±1.77	93.75±1.27	17.83	2.69
+ MHSA	60.27±7.06	61.34±5.07	60.58±4.07	76.54±5.74	93.64±2.69	93.64±1.80	92.56±2.16	92.94±1.66	16.28	2.12
+ DRIFA	63.27±4.48	64.23±4.01	63.25±3.92	77.45±5.17	97.81±2.32	97.81±2.39	94.50±1.60	94.65±0.99	41.91	4.24
+ DVCA	64.04±4.24	65.92±3.50	65.80±2.75	80.14±3.90	98.60±0.27	98.60±0.26	95.61±2.44	97.84±0.64	23.57	2.58
+ EMCAM	64.60±4.66	65.73±3.20	65.80±2.92	81.31±4.29	98.23±1.94	98.30±0.58	95.61±1.92	98.10±2.04	14.42	1.84
+ OTCA	63.95±5.25	64.78±4.21	64.05±3.25	75.22±5.68	94.71±2.30	94.80±0.63	93.20±2.12	93.20±2.50	15.95	2.28
+ CA	64.12±5.49	65.08±4.06	65.17±3.48	78.57±5.08	95.27±2.37	95.27±2.06	94.10±1.87	93.94±1.36	**12.10**	**1.11**
+ HSA	62.85±5.44	63.68±4.52	62.80±3.78	76.25±5.45	94.52±1.88	94.52±2.65	93.75±2.07	93.75±1.47	13.82	1.48
+ HySAM	65.90±4.55	68.05±2.96	66.81±1.62	82.79±4.03	96.60±0.85	97.54±1.14	94.82±0.95	94.50±0.56	14.12	1.59
+ HySAR (Ours)	**69.78**±3.28	**71.87**±2.87	**72.21**±2.03	**86.62**±3.86	**99.10**±0.67	**99.54**±0.52	**96.35**±0.80	**99.10**±0.81	16.47	1.82
MAIL	64.25±4.73	65.39±3.67	65.72±3.06	81.90±4.59	98.60±1.99	98.60±1.19	95.61±2.70	99.27±0.15	11.70	1.84
+ MHSA	60.82±5.64	61.90±4.83	60.23±3.37	75.84±5.41	95.58±0.68	95.80±2.09	93.82±2.41	94.04±2.01	13.95	2.35
+ DRIFA	64.25±5.39	64.62±2.93	64.24±3.31	79.26±4.93	98.23±0.21	98.35±0.73	95.61±0.60	97.84±1.03	42.52	3.13
+ DVCA	65.73±4.09	65.39±3.03	65.47±4.04	81.90±4.45	98.60±0.59	98.60±1.81	95.61±1.15	97.84±2.19	15.71	2.43
+ OTCA	63.91±4.15	64.15±3.08	63.40±3.79	77.25±4.71	96.71±2.46	96.90±1.84	94.29±2.63	94.50±2.14	15.63	2.48
+ ISA	65.07±5.16	66.10±3.51	65.91±3.96	81.31±4.30	97.10±2.15	97.25±0.98	94.82±2.45	94.82±2.42	24.35	2.74
+ CA	63.91±6.04	63.50±4.02	62.24±3.61	78.57±4.53	95.58±2.05	95.58±1.41	94.27±2.56	94.65±2.27	**10.25**	0.63
+ HSA	63.27±4.60	64.83±3.36	63.40±3.49	78.73±5.05	96.90±1.00	96.90±1.74	94.75±1.95	94.75±1.39	10.98	0.68
+ HySAM	65.17±4.18	67.35±2.68	66.57±3.18	82.41±3.19	98.60±0.82	98.60±0.69	95.30±0.75	95.61±0.94	11.15	**0.62**
+ HySAR (Ours)	**70.52**±3.16	**72.18**±2.26	**72.21**±1.97	**85.79**±3.33	**99.07**±0.77	**99.54**±0.42	**96.63**±0.90	**99.81**±0.40	14.46	0.74
CURE	70.39±3.84	72.14±2.40	69.51±3.23	83.28±3.29	98.75±1.26	98.75±0.70	95.97±1.06	99.67±0.09	7.71	0.59
+ MHSA	64.31±6.24	64.79±3.94	63.54±3.93	77.27±5.34	93.85±2.51	93.85±1.82	91.85±2.23	92.17±2.52	10.62	1.55
+ DRIFA	67.10±4.27	67.24±3.73	68.81±3.43	82.37±4.85	98.23±0.25	98.35±0.74	95.61±1.76	97.84±0.97	74.45	7.22
+ DVCA	68.84±4.56	68.70±3.30	69.20±3.35	83.85±4.32	98.30±0.61	98.30±1.08	95.61±1.09	97.84±0.24	14.81	2.32
+ OTCA	66.54±5.15	66.27±4.68	67.35±3.95	80.12±4.80	97.25±2.31	97.25±0.17	94.82±2.48	94.82±1.78	13.27	1.84
+ ISA	69.27±3.91	68.58±3.58	69.80±3.24	82.72±4.31	98.23±2.13	98.30±0.87	95.06±2.43	96.85±2.64	17.82	2.55
+ CA	64.27±4.86	64.31±4.20	63.87±3.65	79.14±5.63	95.90±1.38	95.90±0.86	94.71±2.54	94.80±0.29	**7.25**	**0.54**
+ HSA	65.34±5.51	66.10±3.97	65.52±3.62	80.71±4.10	96.90±0.37	96.90±1.13	94.75±2.10	94.75±1.67	7.53	0.59
+ EMCAM	70.15±4.37	71.26±3.21	69.20±3.76	82.85±4.13	98.23±1.69	98.30±0.54	95.90±1.04	98.33±1.10	9.83	1.27
+ HySAR (Ours)	**74.52**±3.32	**75.18**±2.00	**75.21**±2.02	**88.79**±3.68	**99.54**±0.39	**99.54**±0.51	**96.63**±0.76	**99.81**±0.00	11.82	0.87

2.3 Hybrid-Space Multi-Cross Fusion

Motivation. Although MILA and DUHA learn complementary Euclidean and non-Euclidean evidence, using them independently may underutilize cross-modal complementarity, especially when a modality is noisy, degraded, or missing. We therefore design **Hybrid-Space Multi-Cross Fusion** (HFuse), which cross-fuses target dual-hyperbolic evidence with Euclidean evidence from complementary modalities, thereby balancing Euclidean locality and non-Euclidean hierarchy (see Fig. 3(d)). For each target modality i and complementary modality $r \neq i$, let

$$s_{i\leftarrow r}^{(H,j)} = \alpha_i^j \odot A_i^{(D,j)} + \alpha_r^j \odot A_r^{(E,j)} \tag{12}$$

denote the cross-geometric interaction response. HFuse applies softmax gating over the complementary modalities, followed by a sigmoid $\sigma(\cdot)$ to learn the hybrid-space attention map:

$$A_i^{(H,j)} = \sigma\left(\sum_{\substack{r=1 \\ r\neq i}}^{m} \mathrm{softmax}_r\left(s_{i\leftarrow r}^{(H,j)}\right) \odot s_{i\leftarrow r}^{(H,j)}\right). \tag{13}$$

Here, α_i^j and α_r^j are learnable parameters. The softmax is normalized over the complementary modalities and assigns greater influence to more informative cross-geometric responses, while the sigmoid produces $A_i^{(H,j)}$. With missing modalities, unavailable responses are masked, remaining fusion weights re-normalized, and the learned map yields the robust shared representation in Eq. 3.

3 Experimental Results

Datasets. We evaluate HySAR on ten medical and natural-domain multimodal benchmarks: TCGA BRCA (Lingle et al., 2016), BLCA (Network et al., 2014), UCEC (Erickson et al., 2016), and KIRP (Linehan et al., 2016) using WSI–omics modalities for survival prediction; HAM10000 (Tschandl et al., 2018) and SIPaKMeD (Plissiti et al., 2018) using image–tabular modalities for medical classification; NYUDv2 (Silberman et al., 2012) and SUN-RGB-D (Song et al., 2015) using RGB-D modalities for semantic segmentation; and MM-IMDb (Arevalo et al., 2017) and UPMC Food-101 (Wang et al., 2015) using image–text modalities for classification.

Baselines. For the medical suite, we evaluate HySAR across six host multimodal fusion architectures: MOTCAT (Xu & Chen, 2023), DRIFA-Net (Dhar et al., 2025b), HyPCA-Net (Dhar et al.,

2026a), HEALNet (Hemker et al., 2024), MAIL (Dhar et al., 2025a), and CURE (Dhar et al., 2026b), using ResNet18 (He et al., 2016), ViT-Tiny (Steiner et al., 2022), and Mamba-Tiny (Zhu et al., 2024) backbones. Within each host, its native refiner is independently replaced with Euclidean-space (MHSA (Vaswani et al., 2017), OTCA (Xu & Chen, 2023), DRIFA (Dhar et al., 2025b), DVCA (Dhar et al., 2026a), EMCAM (Dhar et al., 2025a), ISA (Hemker et al., 2024), CA (Chen et al., 2021)), non-Euclidean-manifold (HSA (Hou et al., 2024), HySAM (Dhar et al., 2026b)), or our HySAR refiners.

For the natural-domain suite, we use MANGO (Truong et al., 2026) and HighMMT (Liang et al., 2022) as host multimodal fusion architectures. For RGB-D semantic segmentation, following MANGO (Truong et al., 2026), both RGB and depth modalities use the MAE visual encoder (He et al., 2022) with 256×256 inputs. For classification with image–text modalities, images are encoded using Inception-V3 (Szegedy et al., 2016) with 299×299×3 inputs, while text is encoded using BERT-base (Devlin et al., 2019). Within each host, we compare its native refiner with ISA, DRIFA, EMCAM, HySAM, and HySAR, while keeping the modality encoders, remaining host pipeline, prediction head, and training protocol fixed. **Bold** and underlined denote the best and second-best results, respectively. Results are reported as mean ± standard deviation over 5 random seeds. Dataset and implementation details are available in App. D–E, respectively.

Evaluation Metrics. We report the following metrics: accuracy (ACC), AUC, F1-score, concordance index (C-Index), mean intersection over union (mIoU), *number of parameters* (#P) in millions, *floating-point operations* (#F) in GFLOPs, and *inference time* (#IFT) in milliseconds.

3.1 Performance Comparisons

Medical-vision benchmarks. Table 1 and Figs. 4, 9–11 (see App. F) present results on six multimodal benchmarks (e.g., BRCA, BLCA, UCEC, KIRP HAM10000 and SIPaKMeD) across diverse fusion learners and refiners. The evaluation uses four backbones from three families: ResNet18 and Inception-V3 (convolutional), ViT-Tiny (transformer), and MAMBA-Tiny (state-space). Across these diverse configurations, HySAR-integrated learners improve over their native host configurations by up to ≈ 5.2 points and outperform prior Euclidean-space and non-Euclidean-manifold refiners by up to ≈ 8.4 points, yielding an average overall gain of ≈ 4.7 points. These gains support the effectiveness of HySAR's hybrid-space refinement across diverse backbone inductive biases.

Natural-image benchmarks. Across NYUDv2, SUN-RGB-D, MM-IMDb, and UPMC, HySAR improves MANGO- and HighMMT-based multimodal fusion learners by up to ≈ 7.55 points on average (Table 2). These results demonstrate that HySAR generalizes across heterogeneous modality configurations, host learners, and tasks, highlighting the benefit of hybrid-space refinement. Specifically, MILA preserves fine-grained multi-scale information, DUHA learns complementary hierarchy-aware properties, and HFuse cross-fuses their responses to learn robust shared representations that strengthen multimodal fusion effectiveness.

Table 2: **Performance–efficiency trade-offs on natural-domain benchmarks.** We evaluate plug-and-play cross-modal refiners within MANGO and HighMMT multimodal fusion pipelines on segmentation benchmarks (NYUDv2, SUN-RGB-D) and classification benchmarks (MM-IMDb, UPMC Food-101). Row notation follows Table 1.

Datasets →	NYUDV2	SUN	#P	#F	#IFT	MM-IMDb	UPMC	#P	#F	#IFT
MANGO	59.2	54.1	72.9	152.3	144	71.7	92.7	149.4	75.8	77
+ ISA	58.9	54.5	78.7	170.9	182	71.1	91.6	178.3	86.3	89
+ DRIFA	60.7	55.9	144.7	255.3	305	72.9	93.1	312.5	134.4	154
+ EMCAM	60.5	55.4	48.5	121.8	139	73.5	93.3	145.9	77.9	81
+ HySAM	61.2	55.7	**44.7**	**110.4**	**127**	74.1	93.8	**130.2**	**72.6**	**74**
+ HySAR (Ours)	**65.7**	**61.3**	46.2	115.9	132	**76.4**	**96.5**	135.7	73.1	**74**
HighMMT	52.5	48.3	64.7	123.5	119	65.5	86.6	132.6	71.5	72
+ ISA	55.2	51.7	74.2	159.6	154	67.6	87.4	168.1	84.7	87
+ DRIFA	57.8	53.5	87.3	191.5	192	65.3	85.1	188.5	125.3	145
+ EMCAM	58.2	54.1	41.6	82.5	87	65.5	84.3	125.1	71.7	73
+ HySAM	58.9	54.6	**37.1**	**72.2**	**79**	65.9	85.8	**107.8**	**67.8**	**68**
+ HySAR (Ours)	**63.2**	**59.8**	39.3	78.7	82	**70.7**	**89.4**	115.4	68.5	69

Across the benchmark suite, HySAR's multi-scale information-learning properties and dual hyperbolic manifold representation learning reduce parameters and FLOPs by up to ≈ 84.1% and ≈ 88.0%, respectively. These design choices address **Challenge #1** (***performance–efficiency trade-off***), making HySAR well suited to resource-constrained medical AI. *More results, ablations, and discussion are in App.* F–G.

Missing-modality robustness. Across Omics-only, WSI-only, 50% Both, and 100% Both settings, HySAR improves C-index by ≈ 5.2–8.7 points on BLCA and by up to ≈ 6.3 points on KIRP, with an average gain of up to ≈ 5.1 points across all settings (Fig. 5). These consistent gains across single-, partial-, and complete-modality inference demonstrate that HySAR preserves discriminative

Table 3: **Component ablation of HySAR under a fixed MAIL. (a)** MILA stream ablation on MAIL with DUHA fixed, where MDHC, MGHC, and MDDHC denote depth-wise, group-wise, and dilated depth-wise homogeneous convolution streams, respectively. **(b)** Hyperbolic-attention and hybrid-space design ablations comparing PVRL, PARL, and cascaded (C) versus parallel (P) variants of DUHA and HySAR, with MILA fixed.

(a) MILA stream ablation

MDHC	MGHC	MDDHC	BRCA	BLCA	UCEC	KIRP	#P	#F
×	×	×	66.81	68.32	68.10	82.27	**10.16**	**0.65**
×	×	✓	68.34	68.93	68.81	84.21	11.78	0.68
×	✓	×	67.25	68.54	68.42	83.34	11.63	0.68
✓	×	×	68.72	69.07	68.81	84.50	11.92	0.68
×	✓	✓	68.80	69.42	70.25	84.77	12.74	0.71
✓	×	✓	69.53	71.22	71.08	85.10	12.92	0.71
✓	✓	×	69.17	70.38	70.51	85.10	12.92	0.71
✓	✓	✓	**70.52**	**72.18**	**72.21**	**85.79**	14.46	0.74

(b) Hyperbolic and hybrid manifolds ablation

Approaches	BRCA	BLCA	UCEC	KIRP	#P	#F
PVRL	69.81	71.22	71.08	85.10	**14.35**	**0.74**
PARL	69.20	70.38	70.51	84.77	14.38	**0.74**
DUHA-C	69.53	71.84	71.63	85.37	14.46	**0.74**
DUHA-P	70.52	**72.18**	**72.21**	**85.79**	14.46	**0.74**
HySAR-C	69.90	71.51	71.63	85.10	14.46	**0.74**
HySAR-P	**70.52**	**72.18**	**72.21**	**85.79**	14.46	**0.74**

evidence from the available modalities. Specifically, HySAR mitigates this through hybrid space refinement and availability-aware fusion: MILA retains multi-scale Euclidean cues, DUHA captures complementary hierarchy-aware cues through PARL and PVRL, and HFuse masks unavailable responses before softmax normalization and renormalizes the remaining weights. Consequently, HySAR excludes missing branches while preserving observed-modality representations, thereby addressing **Challenge #2 (*Missing-modality robustness*)** without redesigning the host fusion pipeline.

3.2 Ablation Study

Hyperbolic and Hybrid-space with parallel design. Under the same fixed MAIL host architecture, ResNet18 backbone, and computational cost, Table 3(b) shows that PVRL exceeds PARL by ≈ 0.59 points on average, while parallel DUHA improves over the individual hyperbolic formulations by $\approx$ 0.87–1.46 points. These results confirm that PARL and PVRL encode complementary hierarchy-aware responses. Moreover, parallel DUHA and HySAR outperform their cascaded counterparts by ≈ 0.58 and ≈ 0.64 points, respectively, at identical computational cost. The gain therefore isolates an information-preservation effect within the shared representation refiner: parallel processing exposes both geometric responses to the fusion gate simultaneously, whereas cascading forces the second geometry to operate on an already filtered representation.

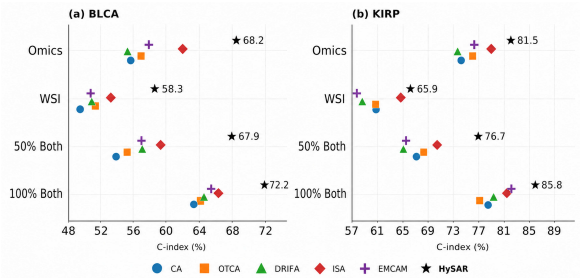

Figure 5: **Missing-modality robustness.** We compare refiners in ResNet18-based multimodal fusion pipelines on BLCA and KIRP under Omics-only, WSI-only, 50% Both, and 100% Both (all modalities available).

Component ablations. Under the MAIL–ResNet18 setting, Table 4 shows that MILA and DUHA improve C-index by ≈ 2.81 and ≈ 2.06 points on average, respectively, while complete HySAR improves it by ≈ 5.86 points on average and reduces FLOPs by $\approx 59.8\%$. Table 3(a) further shows that jointly using MDHC, MGHC, and MDDHC improves C-index by ≈ 3.80 points on average, confirming that the three streams learn complementary evidence[1] rather than redundant Euclidean evidence.

Table 4: Component ablation of HySAR under a fixed MAIL architecture, where MILA, DUHA, and HFuse are progressively added.

Components			BRCA	BLCA	UCEC	KIRP		
MILA	DUHA	HFuse	C-Index	C-Index	C-Index	C-Index	#P	#F
×	×	×	64.25	65.39	65.72	81.90	11.70	1.84
✓	×	×	67.34	68.80	69.21	83.15	14.27	0.74
×	✓	×	66.81	68.32	68.10	82.27	**10.16**	**0.65**
✓	✓	✓	**70.52**	**72.18**	**72.21**	**85.79**	14.46	0.74

4 Conclusion

We introduced HySAR, a plug-and-play hybrid-space shared representation learner that integrates complementary Euclidean and non-Euclidean evidence for efficient and robust multimodal fusion. Across ten medical and natural-domain benchmarks, HySAR-integrated learners outperform prior baselines by ≈ 4.8 points on average, reduce parameters and FLOPs by up to $\approx 84.1\%$ and $\approx 88.0\%$, respectively, and improve missing-modality C-index by up to 8.7 points. These results demonstrate a favorable performance–efficiency–robustness trade-off for resource-constrained multimodal AI. Future work will investigate adversarially robust hybrid-space refinement.

[1] MDHC preserves local-to-coarse spatial evidence, MGHC captures structured spatial responses with intra-group channel interactions, and MDDHC enlarges the effective receptive field for broader contextual aggregation.

AI USE STATEMENT

We selected few sentences at a time and asked GPT to check the grammar.

REFERENCES

Yajun An, Jiale Chen, Huan Lin, Zhenbing Liu, Siyang Feng, Hualong Zhang, Rushi Lan, Zaiyi Liu, and Xipeng Pan. CA-MLIF: Cross-attention and multimodal low-rank interaction fusion framework for tumor prognostic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1764–1772, 2025.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 357–366, 2021.

Ziheng Chen, Zihan Su, Bernhard Schölkopf, and Nicu Sebe. Proper velocity neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 122721–122760, 2026.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35 (1):126–136, 2018.

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

Joy Dhar, Puneet Goyal, Maryam Haghighat, Nayyar Zaidi, Ferdous Sohel, Bao Q Vo, and KC Santosh. Towards building robust models for unimodal and multimodal medical imaging data. *Information fusion*, pp. 103822, 2025a.

Joy Dhar, Nayyar Zaidi, Maryam Haghighat, Sudipta Roy, Puneet Goyal, Azadeh Alavi, and Vikas Kumar. Multimodal fusion learning with dual attention for medical imaging (DRIFA-Net). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025b.

Joy Dhar, Manish Kumar Pandey, Debashis Das Chakladar, Maryam Haghighat, Azadeh Alavi, Sajib Mistry, and Nayyar Zaidi. Hypca-net: Advancing multimodal fusion in medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1831–1840, 2026a.

Joy Dhar, Manish Kumar Pandey, Nayyar Zaidi, Chen Chen, Maryam Haghighat, Ferdous Sohel, and Puneet Goyal. Advancing multimodal fusion on heterogeneous medical data with hybrid geometry attention. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 863–874, 2026b.

Joy Dhar, Nayyar Zaidi, and Maryam Haghighat. Effective and robust multimodal medical image analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 188–199, 2026c.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11): 3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

Bradley J. Erickson, David Mutch, Lynne Lippmann, and Rose Jarosz. The cancer genome atlas uterine corpus endometrial carcinoma collection (tcga-ucec), 2016. URL https://www.cancerimagingarchive.net/collection/tcga-ucec/.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Caglar Gulcehre, Misha Denil, Mateusz Malinowski, Ali Razavi, Razvan Pascanu, Karl Moritz Hermann, Peter Battaglia, Victor Bapst, David Raposo, Adam Santoro, et al. Hyperbolic attention networks. *arXiv preprint arXiv:1805.09786*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. HEALNet: Multimodal fusion for heterogeneous biomedical data. *Advances in Neural Information Processing Systems*, 37:64479–64498, 2024.

Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Multimodal lego: Model merging and fine-tuning across topologies and modalities in biomedicine. In *International Conference on Learning Representations*, volume 2025, pp. 51822–51843, 2025.

Peizhong Hou, Haiyang Wang, Tianming Li, and Junchi Yan. HSA: Hyperbolic self-attention for sequential recommendation. In *Web and Big Data (APWeb-WAIM 2023)*, volume 14333 of *Lecture Notes in Computer Science*, pp. 250–264. 2024. doi: 10.1007/978-981-97-2387-4_17.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13289–13299, 2020.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

Marston Linehan, Rabindra Gautam, Shanah Kirk, Yueh Lee, Charles Roche, Ermelinda Bonaccio, Joe Filippini, Kimberly Rieger-Christ, John Lemmerman, and Rose Jarosz. The cancer genome atlas cervical kidney renal papillary cell carcinoma collection (tcga-kirp), 2016. URL https://www.cancerimagingarchive.net/collection/tcga-kirp/.

W Lingle, BJ Erickson, ML Zuley, R Jarosz, E Bonaccio, J Filippini, and N Gruszauskas. Radiology data from the cancer genome atlas breast invasive carcinoma [TCGA-BRCA] collection. *The Cancer Imaging Archive*, 10(K9):5, 2016.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315, 2014.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems 30*, pp. 6341–6350, 2017. doi: 10.5555/3295222.3295381.

Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(12):10023–10044, 2021.

Maria E Plissiti, Panagiotis Dimitrakopoulos, Giorgos Sfikas, Christophoros Nikou, Orestis Krikoni, and Avraam Charchanti. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3144–3148. IEEE, October 2018.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pp. 746–760. Springer, 2012.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. Multimodn—multimodal, multi-task, interpretable modular networks. *Advances in neural information processing systems*, 36:28115–28138, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Thanh-Dat Truong, Christophe Bobda, Nitin Agarwal, and Khoa Luu. Mango: Multimodal attention-based normalizing flow approach to fusion learning. *Advances in Neural Information Processing Systems*, 38:138213–138234, 2026.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1): 1–9, 2018.

Abraham Albert Ungar. *Analytic hyperbolic geometry and Albert Einstein's special theory of relativity*. World Scientific, 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International conference on multimedia & expo workshops (ICMEW)*, pp. 1–6. IEEE, 2015.

Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21241–21251, 2023.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 2024.

APPENDIX

A RELATED STUDY

Cross-modal refinement within multimodal fusion architectures. Multimodal fusion architectures determine when and where heterogeneous modality features interact, whereas their cross-modal refiners determine how encoded modality evidence is transformed into shared representations. For example, MOTCAT (Xu & Chen, 2023), DRIFA-Net (Dhar et al., 2025b), HyPCA-Net (Dhar et al., 2026a), HEALNet (Hemker et al., 2024), MAIL (Dhar et al., 2025a), and CURE (Dhar et al., 2026b) define different host fusion pipelines, while their internal refiners govern shared representation learning. This distinction is important because improvements obtained by replacing only the refiner can be attributed to representation refinement rather than to a redesign of the modality encoders, backbone, prediction head, or overall fusion topology.

Euclidean-space cross-modal refiners. MHSA (Vaswani et al., 2017), OTCA (Xu & Chen, 2023), DRIFA (Dhar et al., 2025b), DVCA (Dhar et al., 2026a), EMCAM (Dhar et al., 2025a), ISA (Hemker et al., 2024), and CA (Chen et al., 2021) are all cross-modal refiners that operate within Euclidean space. Although they employ different attention, alignment, and feature-calibration operators, they share the objective of capturing fine-grained dependencies among encoded modality features. Their principal strength lies in preserving spatial, channel, or latent interaction evidence. Their common limitation is therefore not an absence of cross-modal interaction; rather, the interactions remain represented within a single Euclidean geometric family, which is less suited to explicitly preserving coarse-to-fine hierarchy-aware structural relations. Moreover, several such refiners rely on high-dimensional affinity operations or repeated refinement, creating a tension between representation expressiveness and computational cost.

Non-Euclidean-manifold cross-modal refiners. Non-Euclidean cross-modal refiners, including HSA (Hou et al., 2024) and HySAM (Dhar et al., 2026b), introduce geometry-aware inductive biases for capturing hierarchy-aware structural dependencies. These approaches improve the representation of relations whose organization is difficult to preserve using Euclidean refinement alone. Nevertheless, prioritizing non-Euclidean-manifold responses does not explicitly retain complementary multi-scale Euclidean spatial–channel evidence within the same refiner. Consequently, Euclidean and non-Euclidean refiners emphasize complementary forms of evidence, but learning either geometric family in isolation limits shared-representation expressiveness.

Efficient and missing-modality-aware refinement. Cross-modal refinement must preserve informative evidence when only a subset of modalities is available, without introducing prohibitive computational overhead. Many fusion pipelines are optimized for complete inputs and delegate missing-modality handling to external masking, imputation, or a subsequent fusion stage. Such treatment may neither preserve modality-local evidence nor exclude unavailable responses before cross-modal normalization. Conversely, increasing refiner expressiveness through dense attention or additional processing blocks raises parameter and computational costs. These limitations motivate a lightweight refiner that performs availability-aware interaction while preserving evidence from the available modalities.

Refiner-level design-space synthesis. Table 5 synthesizes the geometric and missing-modality properties of representative cross-modal refiners. Euclidean-space refiners primarily preserve fine-grained spatial, channel, token, or latent dependencies, whereas HSA and HySAM emphasize hierarchy-aware non-Euclidean structure. However, the compared approaches generally specialize in an individual geometric family and do not directly coordinate fine-grained Euclidean evidence with hierarchy-aware hyperbolic evidence within the same refiner. Moreover, missing-modality handling is commonly absent at the refiner level or delegated to the corresponding host architecture.

Research gap. The unresolved question is not whether existing refiners model cross-modal interactions—they do—but whether a single plug-and-play refiner can jointly **(i)** preserve fine-grained, multi-scale Euclidean evidence; **(ii)** capture diversified hierarchy-aware responses through complementary formulations of hyperbolic space; **(iii)** cross-fuse Euclidean and dual-hyperbolic evidence across only the available modalities; and **(iv)** remain computationally efficient and robust under missing inputs. Among the compared refiners, this unified representation-level capability remains underexplored.

Table 5: **Conceptual comparison of representative cross-modal refiners.** We compare their representation formulation, explicitly modeled dependencies, direct Euclidean–hyperbolic fusion, and missing-modality handling.

Cross-modal refiner	Representation formulation	Fine-grained dependencies	Hierarchy-aware dependencies	Euclidean–hyperbolic fusion	Missing-modality handling
Euclidean-space refiners					
MHSA	Euclidean	**Yes**	–	–	–
OTCA	Euclidean	**Yes**	–	–	–
DRIFA	Euclidean	**Yes**	–	–	–
DVCA	Euclidean	**Yes**	–	–	–
EMCAM	Euclidean	**Yes**	–	–	–
ISA	Euclidean	**Yes**	–	–	Host-level
CA	Euclidean	**Yes**	–	–	–
Hyperbolic-based refiners					
HSA	Hyperbolic	–	**Yes**	–	–
HySAM	Hyperbolic + quantum-inspired	–	**Yes**	–	Host-level
Hybrid Euclidean–hyperbolic refiner					
HySAR (Ours)	Euclidean + Hyperbolic	**Yes**	**Yes**	**Yes**	**Refiner-level**

Interpretation: Fine-grained dependencies comprise spatial, channel, token, or latent cross-modal relations, whereas hierarchy-aware dependencies denote explicit hyperbolic modeling of coarse-to-fine structural relations. Poincaré, Lorentz, and Proper-Velocity are distinct formulations of hyperbolic space. Euclidean–hyperbolic fusion denotes direct integration of both responses within the same cross-modal refiner. "Host-level" indicates that missing-modality handling is provided by the surrounding host architecture, whereas "refiner-level" denotes explicit availability-aware masking and renormalization within the refiner itself. "–" indicates no explicit mechanism.

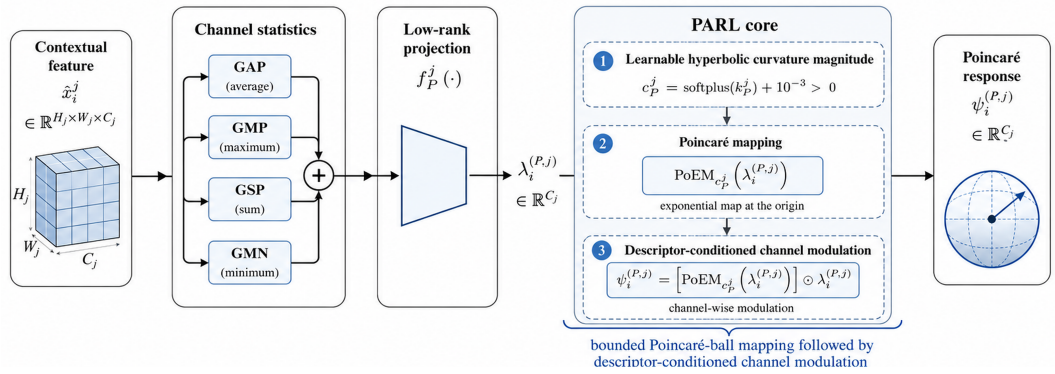

Figure 6: **Poincaré-Aware Representation Learning (PARL).** The contextual feature $\hat{x}_i^j$ is summarized using complementary channel statistics (GAP, GMP, GSP, and GMN) and a low-rank projection to obtain $\lambda_i^{(P,j)}$. PARL applies an origin-centered Poincaré exponential map with learnable curvature magnitude c_P^j, followed by descriptor-conditioned channel modulation, to learn the hierarchy-aware Poincaré response $\psi_i^{(P,j)}$.

Our approach. HySAR addresses this gap without redesigning the host multimodal fusion architecture. MILA preserves channel-local, group-structured, and context-expanded Euclidean evidence through lightweight multi-scale homogeneous convolution streams. DUHA captures complementary hierarchy-aware responses through parallel PARL and PVRL, which instantiate the bounded Poincaré-ball and unconstrained Proper-Velocity formulations of hyperbolic space, respectively. Availability-aware HFuse then adaptively cross-fuses the Euclidean and dual-hyperbolic responses over the available modalities, masks unavailable responses before softmax normalization, and renormalizes the remaining valid weights. Thus, HySAR advances plug-and-play cross-modal refinement beyond isolated geometric families toward efficient, availability-aware hybrid Euclidean–hyperbolic shared representation learning.

B Additional Method Details

This section provides additional details of the two complementary non-Euclidean representation learners within DUHA: Poincaré-Aware Representation Learning (PARL) and Proper-Velocity Manifold Representation Learning (PVRL). Their internal structures are illustrated in Figs. 6 and 7, respectively.

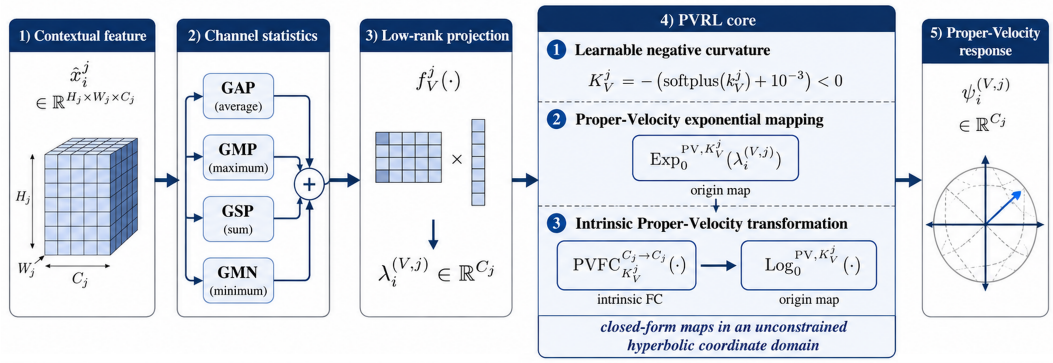

Figure 7: **Proper-Velocity Manifold Representation Learning (PVRL).** The contextual feature $\hat{x}_i^j$ is summarized using complementary channel statistics (GAP, GMP, GSP, and GMN) and a low-rank projection to obtain $\lambda_i^{(V,j)}$. PVRL maps this descriptor through the origin-centered PV exponential map with learnable negative curvature K_V^j, applies an intrinsic PVFC transformation, and returns through the corresponding logarithmic map to learn the hierarchy-aware response $\psi_i^{(V,j)}$.

Shared descriptor construction. Given the contextual feature $\hat{x}_i^j \in \mathbb{R}^{H_j \times W_j \times C_j}$, both branches first summarize complementary channel statistics through GAP, GMP, GSP, and GMN. Their aggregation captures average contextual evidence, salient responses, total response magnitude, and complementary low-response structure. Independent low-rank projections $f_P^j(\cdot)$ and $f_V^j(\cdot)$ then produce the geometry-specific descriptors $\lambda_i^{(P,j)}$ and $\lambda_i^{(V,j)}$, respectively. This shared descriptor construction provides both branches with the same contextual evidence while allowing their subsequent geometric transformations to remain formulation-specific.

Poincaré-aware representation learning. Figure 6 details PARL. Given $\lambda_i^{(P,j)}$, PARL learns a curvature-aware hierarchical response on the Poincaré formulation of hyperbolic space. A learnable curvature magnitude $c_P^j = \mathrm{softplus}(k_P^j) + 10^{-3}$ parameterizes the origin-centered Poincaré exponential mapping. The resulting hyperbolic embedding is subsequently combined with the original descriptor through channel-wise modulation, yielding the Poincaré-aware response $\psi_i^{(P,j)}$ in Eq. 10. This construction preserves descriptor-specific response magnitudes while injecting curvature-dependent radial hierarchy into the channel representation.

Proper-Velocity manifold representation learning. Figure 7 illustrates PVRL. Unlike the bounded Poincaré formulation, PVRL exploits the unconstrained coordinate representation of the Proper-Velocity (PV) manifold. Given $\lambda_i^{(V,j)}$, a learnable negative curvature $K_V^j = -(\mathrm{softplus}(k_V^j) + 10^{-3})$ parameterizes the origin-centered PV exponential map. The mapped descriptor is transformed through the intrinsic $\mathrm{PVFC}_{K_V^j}^{C_j \to C_j}(\cdot)$ operation and returned through the corresponding PV logarithmic map, producing the PV-aware response $\psi_i^{(V,j)}$ in Eq. 11. Thus, PVRL models hierarchy-aware channel transformations within an unconstrained hyperbolic coordinate domain.

Complementary dual-hyperbolic refinement. The two branches encode complementary geometric biases: PARL emphasizes bounded, curvature-aware radial structure, whereas PVRL provides hierarchy-aware transformations through an unconstrained PV coordinate representation. Rather than cascading these transformations, DUHA evaluates them in parallel from independently projected descriptors and adaptively combines their responses through geometry-wise softmax gating followed by sigmoid calibration, as defined in Eq. 9. The resulting dual-hyperbolic attention map $A_i^{(D,j)}$ therefore retains complementary hierarchy-aware evidence before its subsequent cross-geometric integration with the Euclidean response in HFuse.

C Algorithms

This section provides the pseudo-code of HySAR (Algorithm 1) and its components. Let $\mathcal{O} \subseteq \{1, \ldots, m\}$ denote the set of modalities available at inference. At each block j, HySAR first invokes Algorithm 2 and Algorithm 3 for every $i \in \mathcal{O}$ to obtain the Euclidean and dual-hyperbolic attention

Algorithm 1 HySAR shared representation refinement

Require: Encoded features $\{\hat{x}_i^j\}_{i\in\mathcal{O}}$ at block j; available-modality set $\mathcal{O}$; learnable modulations $\{\alpha_i^j\}_{i\in\mathcal{O}}$
Ensure: Shared representations $\{x_i^{S,j}\}_{i\in\mathcal{O}}$
for all $i \in \mathcal{O}$ **do**
  $A_i^{(E,j)} \leftarrow \mathrm{MILA}(\hat{x}_i^j)$
  $A_i^{(D,j)} \leftarrow \mathrm{DUHA}(\hat{x}_i^j)$
end for
for all $i \in \mathcal{O}$ **do**
  $A_i^{(H,j)} \leftarrow \mathrm{HFUSE}(A_i^{(D,j)}, A_r^{(E,j)})$
  $x_i^{S,j} \leftarrow \hat{x}_i^j \odot A_i^{(H,j)} \odot \alpha_i^j$
end for
return $\{x_i^{S,j}\}_{i\in\mathcal{O}}$

Algorithm 2 MILA: multi-scale Euclidean attention

Require: Encoded feature $\hat{x}_i^j$; stages $p \in \{1, 2\}$
Ensure: Euclidean attention map $A_i^{(E,j)}$
$h_i^{(1,j)} \leftarrow \hat{x}_i^j$
for $p = 1$ to 2 **do**
  $\varsigma_{*,i}^{(p,j)} \leftarrow \mathrm{MDHC}(h_i^{(p,j)})$
  $\lambda_i^{(p,j)} \leftarrow \mathrm{MGHC}(\varsigma_{*,i}^{(p,j)})$
  $\xi_i^{(p,j)} \leftarrow \mathrm{MDDHC}(\varsigma_{*,i}^{(p,j)})$
  $h_i^{(p+1,j)} \leftarrow \mathcal{I}\Big(\varsigma_{*,i}^{(p,j)}, \lambda_i^{(p,j)}, \xi_i^{(p,j)}\Big)$
  $\zeta_i^{(p,j)} \leftarrow \mathcal{I}\Big(\mathrm{GMP}(\varsigma_{*,i}^{(p,j)}), \mathrm{GAP}(\lambda_i^{(p,j)}), \mathrm{GMN}(\xi_i^{(p,j)})\Big)$
end for
$z_i^j \leftarrow \mathrm{CONCAT}(\zeta_i^{(1,j)}, \zeta_i^{(2,j)})$
$A_i^{(E,j)} \leftarrow \sigma\Big(\mathrm{LRD}(z_i^j)\Big)$
return $A_i^{(E,j)}$

maps, respectively. Algorithm 4 then cross-fuses these responses over the available complementary modalities to obtain the hybrid-space attention map. The complete calling sequence is

$$\begin{aligned} A_i^{(E,j)} &\leftarrow \texttt{MILA}\Big(\hat{x}_i^j\Big), \\ A_i^{(D,j)} &\leftarrow \texttt{DUHA}\Big(\hat{x}_i^j\Big), \\ A_i^{(H,j)} &\leftarrow \texttt{HFuse}_i^j\Big(A_i^{(D,j)}, \Big\{A_r^{(E,j)}\Big\}_{r\neq i}\Big), \\ x_i^{S,j} &\leftarrow \hat{x}_i^j \odot A_i^{(H,j)} \odot \alpha_i^j. \end{aligned} \tag{14}$$

The resulting $x_i^{S,j}$ is forwarded to the subsequent fusion block when $j < J$ and to the task-specific prediction head when $j = J$.

D Additional Dataset Details

Table 6 briefly summarizes all datasets used in this study.

D.1 Dataset Preprocessing Details

Medical WSI–omics benchmarks. For BRCA, BLCA, UCEC, and KIRP, we retain patient-matched WSI, omics, survival-time, and censoring records, with all records from the same patient assigned

Algorithm 3 DUHA: dual-hyperbolic attention

Require: Encoded feature $\hat{x}_i^j$; learnable channel modulations $\omega_i^{(P,j)}$, $\omega_i^{(V,j)}$
Ensure: Dual-hyperbolic attention map $A_i^{(D,j)}$

$d_i^j \leftarrow \mathrm{GAP}(\hat{x}_i^j) + \mathrm{GMP}(\hat{x}_i^j) + \mathrm{GSP}(\hat{x}_i^j) + \mathrm{GMN}(\hat{x}_i^j)$
$\lambda_i^{(P,j)} \leftarrow f_P^j(d_i^j)$; $\lambda_i^{(V,j)} \leftarrow f_V^j(d_i^j)$
$\psi_i^{(P,j)} \leftarrow \mathrm{PARL}(\lambda_i^{(P,j)})$
$\psi_i^{(V,j)} \leftarrow \mathrm{PVRL}(\lambda_i^{(V,j)})$
$u_i^{(P,j)} \leftarrow \omega_i^{(P,j)} \odot \psi_i^{(P,j)}$
$u_i^{(V,j)} \leftarrow \omega_i^{(V,j)} \odot \psi_i^{(V,j)}$
$(\beta_i^{(P,j)}, \beta_i^{(V,j)}) \leftarrow \textsc{Softmax}(u_i^{(P,j)}, u_i^{(V,j)})$
$A_i^{(D,j)} \leftarrow \sigma\Big(\beta_i^{(P,j)} \odot u_i^{(P,j)} + \beta_i^{(V,j)} \odot u_i^{(V,j)}\Big)$
return $A_i^{(D,j)}$

Algorithm 4 HFuse: missing-modality-aware cross-geometric fusion

Require: Target map $A_i^{(D,j)}$; Euclidean maps $\{A_r^{(E,j)}\}_{r\in\mathcal{O}\setminus\{i\}}$; available set $\mathcal{O}$; learnable modulations $\{\alpha_r^j\}_{r\in\mathcal{O}}$
Ensure: Hybrid-space attention map $A_i^{(H,j)}$

$\mathcal{S}_i \leftarrow \mathcal{O} \setminus \{i\}$
if $\mathcal{S}_i = \varnothing$ **then**
    $A_i^{(H,j)} \leftarrow \sigma\Big(\alpha_i^j \odot A_i^{(D,j)}\Big)$ ▷ single-modality fallback; must match the released code
else
    for all $r \in \mathcal{S}_i$ **do**
        $s_{i\leftarrow r}^{(H,j)} \leftarrow \alpha_i^j \odot A_i^{(D,j)} + \alpha_r^j \odot A_r^{(E,j)}$
    end for
    $\{\gamma_{i\leftarrow r}^{(j)}\}_{r\in\mathcal{S}_i} \leftarrow \textsc{Softmax}(\{s_{i\leftarrow r}^{(H,j)}\}_{r\in\mathcal{S}_i})$
    $A_i^{(H,j)} \leftarrow \sigma\Big(\sum_{r\in\mathcal{S}_i} \gamma_{i\leftarrow r}^{(j)} \odot s_{i\leftarrow r}^{(H,j)}\Big)$
end if
return $A_i^{(H,j)}$

to a single partition. Each native WSI, $X_i \in \mathbb{R}^{H_i\times W_i\times 3}$, is processed using TRIDENT with its HEST tissue segmenter (Zhang et al., 2025) to identify tissue-containing regions and their spatial coordinates. Following the whole-slide preprocessing configuration of TITAN (Ding et al., 2025), non-overlapping $512 \times 512 \times 3$ tissue patches are extracted at $20\times$ magnification and encoded by the pathology-specific CONCHv1.5 encoder (Lu et al., 2024; Ding et al., 2025), yielding a 768-dimensional representation per patch. A slide containing N_i retained patches is therefore represented as $F_i \in \mathbb{R}^{N_i\times 768}$ together with the normalized two-dimensional coordinates of its constituent patches.

Following the structured-query formulation of Perceiver IO (Jaegle et al., 2022), the spatially indexed patch representations are mapped through coordinate-aware cross-attention to 16,384 learnable output tokens arranged on a 128×128 grid, yielding $F_i^s \in \mathbb{R}^{128\times 128\times 768}$. Successive 1×1 channel projections then reduce the feature dimension as $768 \to 128 \to 32 \to 8 \to 3$, while retaining the 128×128 spatial organization, producing the compact slide representation $Z_i^s \in \mathbb{R}^{128\times 128\times 3}$ supplied to the multimodal fusion backbone.

Following CURE (Dhar et al., 2026b), each non-imaging omics sample is initially represented as a D_i-dimensional vector $x_i^o \in \mathbb{R}^{D_i}$ and reshaped into a pseudo-image tensor $x_i^o \in \mathbb{R}^{H_i\times W_i\times C_i}$ for uniform convolutional processing across heterogeneous modalities. In our setting, this representation is standardized to $x_i^o \in \mathbb{R}^{128\times 128\times 3}$, matching the spatial and channel dimensionality of the compact WSI representation before multimodal fusion. We use $80\%/10\%/10\%$ training/validation/test partitions with strict patient-level separation.

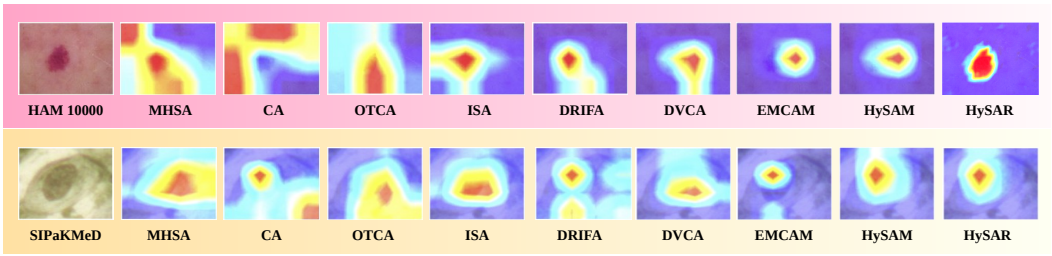

Figure 8: **Grad-CAM comparison of cross-modal refiners within the CURE-based multimodal fusion pipeline.** Heatmaps visualize regions contributing positively to the selected prediction target, enabling qualitative comparison of prediction-related spatial emphasis between HySAR and baseline refiners.

Medical image–tabular benchmarks. For HAM10000 and SIPaKMeD, we retain their multimodal image–tabular formulation rather than treating them as imaging-only benchmarks. Following the medical preprocessing protocol of CURE (Dhar et al., 2026b), images are resized to $128 \times 128 \times 3$, while the corresponding tabular representations are transformed to the same input dimensionality and paired with their associated images. We use $80\%/10\%/10\%$ training/validation/test partitions, keeping paired records within the same split.

Medical-data augmentation. Following CURE (Dhar et al., 2026b), each training image in the image–tabular benchmarks is augmented with three label-preserving variants: a 20° rotation, a 5-pixel horizontal/vertical translation, and Gaussian smoothing with a 3×3 kernel. Augmented images retain their corresponding labels and paired tabular information. Augmentation is restricted to the training partition.

RGB-D semantic segmentation. Following MANGO (Truong et al., 2026), both RGB and depth modalities of NYUDv2 and SUN-RGB-D use the MAE visual encoder (He et al., 2022) with 256×256 inputs. We retain the corresponding $795/654$ and $5{,}285/5{,}050$ training/test partitions, respectively, together with the task-specific segmentation preprocessing and prediction heads.

Image–text classification. For MM-IMDb, we retain the MANGO (Truong et al., 2026) data partition, using 15,552 training and 2,608 validation samples. Images are processed by Inception-V3 (Szegedy et al., 2016) at $299\times299\times3$, while text is encoded by BERT-base (Devlin et al., 2019). UPMC Food-101 uses the same image–text preprocessing and modality encoders while retaining its dataset-specific partition.

E Implementation and Training Details

Backbones and modality encoders. For medical experiments, we evaluate the host multimodal fusion pipelines with ResNet18 (He et al., 2016), ViT-Tiny (Steiner et al., 2022), Mamba-Tiny (Zhu et al., 2024), and Inception-V3 (Szegedy et al., 2016) in the corresponding cross-backbone comparisons. The modality representations supplied to these backbones follow Sec. D.1. For natural-domain experiments, the modality encoders retain the task-specific configurations of MANGO (Truong et al., 2026).

Optimization protocol and task-specific losses. All medical downstream models are trained for 200 epochs on a single NVIDIA A100 80 GB GPU under Ubuntu. We use negative log-likelihood (NLL) for WSI–omics survival prediction and cross-entropy for image–tabular classification. Optimization uses Adam with an initial learning rate of 1×10^{-3} and ReduceLROnPlateau with a minimum learning rate of 1×10^{-6}. For learning the compact WSI representation defined in Sec. D.1, we additionally optimize

$$\mathcal{L}_{\text{comp}} = \lambda_{\text{rec}} \frac{\left\| F_i^s - \widehat{F}_i^s \right\|_1}{128 \times 128 \times 768} + \lambda_{\cos} \left[1 - \cos\left(F_i^s, \widehat{F}_i^s \right) \right],$$

where $\widehat{F}_i^s$ is reconstructed from the $128 \times 128 \times 3$ compact representation during compressor training. Natural-domain experiments retain the task-specific training hyperparameters and prediction objectives of MANGO (Truong et al., 2026).

Table 6: Composition of the ten multimodal benchmarks. For TCGA cohorts, N counts complete patient-level WSI–omics pairs.

Dataset	Modalities	Task	cohort
Medical vision benchmarks			
BRCA	WSI and omics	Survival prediction	$N = 1{,}019$ samples
BLCA	WSI and omics	Survival prediction	$N = 436$ samples
UCEC	WSI and omics	Survival prediction	$N = 538$ samples
KIRP	WSI and omics	Survival prediction	$N = 284$ samples
HAM10000	Dermoscopy image and clinical data	7-class skin-lesion classification	10,015 samples
SIPaKMeD	Cervical-cell image and clinical data	5-class cervical-cell classification	4,049 samples
Natural image benchmarks			
NYUDv2	RGB and depth	40-class semantic segmentation	1,449 samples
SUN-RGB-D	RGB and depth	37-class semantic segmentation	10,335 labeled pairs
MM-IMDb	Image and text	Multi-label classification over 23 genres	25,959 samples
UPMC Food-101	Image and text	101-class food classification	90,840 samples

Replacement protocol. All dataset preprocessing and modality representations are fixed across the corresponding comparisons. Within each host–dataset setting, only the cross-modal shared representation refiner is replaced; the modality encoders, backbone, remaining host pipeline, prediction head, and training protocol remain unchanged. Results are reported as mean $\pm$ standard deviation over five random seeds, isolating the contribution of HySAR.

F Additional Experimental Results and Ablation Studies

Qualitative analysis. Figure 8 presents a qualitative comparison of diverse cross-modal refiners integrated into the CURE-based multimodal fusion pipeline, using Grad-CAM to examine how HySAR and baseline refiners influence prediction-related spatial emphasis. Euclidean-only refiners emphasize spatial–channel dependencies, whereas CURE's native HySAM emphasizes non-Euclidean representation learning. HySAR explicitly integrates multi-scale Euclidean evidence with dual-hyperbolic channel responses, providing complementary information for refining the representations used by CURE's downstream predictor.

Specifically, MILA derives channel attention from multi-scale convolutional responses, allowing refinement to account for spatial patterns at different scales. DUHA complements this evidence through parallel PARL and PVRL transformations of channel descriptors. Their learned combination provides geometry-aware calibration of modality-specific responses. HFuse subsequently combines the target modality's dual-hyperbolic response with Euclidean responses from complementary modalities. The resulting channel recalibration therefore incorporates both geometric representations and cross-modal evidence.

This coordinated refinement can favor channels carrying task-relevant information and reduce the relative influence of less informative responses. Because individual channels encode different spatial patterns, their recalibration can change the spatial evidence contributing to the final prediction. This provides a mechanism through which HySAR can support more selective use of predictive features than single-geometry refiners within CURE. Grad-CAM offers a qualitative view of the resulting spatial emphasis, while attributing specific improvements to individual components requires separate ablation evidence.

F.1 Cross-Backbone Generalization

Table 7 and Fig. 9 provide the performance comparison with the ResNet18 backbone across the six medical benchmarks and diverse host multimodal fusion learners. Figures 10 and 11 extend

Table 7: **Performance–efficiency trade-offs across additional host pipelines.** We evaluate plug-and-play cross-modal refiners within ResNet18-based DRIFA-Net, MOTCAT, and HyPCA-Net multimodal fusion pipelines across six medical benchmarks: BRCA, BLCA, UCEC, KIRP, HAM10000, and SIPaKMeD. Row notation follows Table 1.

Datasets →	BRCA	BLCA	UCEC	KIRP	HAM10000		SIPaKMeD		#P	#F
DRIFA-Net	62.27±5.10	63.74±3.87	62.58±3.07	78.92±5.13	97.81±1.73	97.81±1.11	95.30±1.72	95.27±1.24	53.84	4.83
+ MHSA	60.52±7.50	60.64±5.11	59.78±3.45	75.31±4.87	93.85±1.70	94.10±0.20	92.63±1.45	92.80±1.59	30.58	3.61
+ OTCA	63.27±4.50	64.74±3.40	61.25±2.82	78.15±4.34	95.61±1.71	95.78±0.19	93.82±1.93	94.04±1.33	33.75	3.13
+ DVCA	64.31±3.74	65.27±2.84	65.10±3.60	81.31±4.36	98.60±2.67	98.60±2.17	95.61±2.62	97.84±1.16	34.70	3.22
+ EMCAM	63.91±4.08	64.85±4.38	64.25±2.89	80.45±5.53	98.23±1.23	98.35±2.21	95.61±1.52	97.84±1.25	26.72	2.48
+ ISA	64.90±4.88	66.79±4.28	65.38±2.86	81.90±4.54	96.81±1.96	96.90±1.68	94.27±1.12	94.65±1.90	35.10	3.58
+ CA	63.95±6.02	63.18±3.72	62.40±3.98	79.54±5.09	94.52±0.36	94.41±1.46	93.82±1.98	93.82±1.35	**24.92**	**1.95**
+ HSA	61.38±6.38	62.05±3.64	61.85±3.04	78.92±4.33	93.64±1.37	93.64±1.57	93.75±1.86	93.75±2.20	26.38	2.08
+ HySAM	65.17±4.58	67.58±3.66	66.31±2.90	82.25±3.99	98.60±1.05	98.60±0.96	95.30±1.32	97.63±1.18	26.74	2.21
+ HySAR (Ours)	**68.84**±2.99	**70.27**±1.79	**70.50**±2.77	**84.08**±3.26	**99.10**±0.84	**99.25**±0.44	**96.63**±0.72	**99.50**±0.53	31.54	2.53
MOTCAT	63.73±4.62	64.34±3.22	63.80±4.51	79.59±5.37	93.35±1.75	93.48±0.92	92.56±1.58	92.94±2.35	16.54	1.48
+ MHSA	60.27±5.23	60.64±3.63	61.25±4.05	75.38±5.72	92.91±2.55	92.91±1.55	91.85±2.66	92.17±1.53	15.30	2.12
+ DRIFA	62.91±5.02	63.82±4.16	62.73±2.71	77.14±4.69	97.10±2.33	97.25±0.88	94.27±1.44	94.30±0.18	39.58	3.91
+ DVCA	63.27±4.23	64.56±3.42	64.63±2.91	80.42±3.81	97.54±0.32	97.40±1.21	94.50±1.49	94.50±1.28	20.84	2.45
+ EMCAM	63.10±3.69	64.23±2.88	64.10±2.85	79.90±4.22	96.84±1.54	97.15±0.48	94.27±0.30	94.50±0.55	14.72	1.84
+ ISA	64.52±4.17	66.47±3.01	65.10±2.72	81.90±4.35	95.58±1.83	95.27±1.42	93.91±1.61	93.75±2.34	24.84	2.87
+ CA	63.73±5.06	64.79±4.11	64.25±2.83	80.42±4.39	93.64±2.29	93.90±0.89	93.23±2.11	92.88±2.61	**12.75**	**1.12**
+ HSA	61.91±5.69	61.58±4.47	62.24±3.47	76.21±4.44	94.41±1.50	94.80±1.30	93.08±1.40	92.56±2.38	13.47	1.29
+ HySAM	64.38±3.70	67.20±2.58	66.16±2.79	82.57±4.25	97.54±1.02	97.54±0.49	94.82±0.71	94.50±0.83	13.62	1.41
+ HySAR (Ours)	**68.52**±2.18	**70.27**±3.54	**69.38**±2.28	**85.31**±2.25	**98.85**±0.57	**99.07**±0.45	**95.90**±0.38	**97.84**±0.35	15.25	1.84
HyPCA-Net	64.54±5.66	65.85±3.80	64.73±3.34	80.14±3.75	98.60±0.33	98.60±0.47	95.70±0.31	98.12±1.89	14.47	2.25
+ MHSA	61.10±5.01	62.27±4.61	60.54±4.26	76.24±5.85	94.52±1.17	94.52±1.56	93.08±2.24	92.80±2.22	11.75	2.08
+ OTCA	63.91±5.75	64.17±4.63	62.38±2.94	77.25±4.49	96.42±2.47	96.80±1.22	94.29±2.59	94.50±2.36	13.27	2.14
+ DRIFA	63.75±4.72	65.21±3.15	64.52±3.27	78.92±4.57	98.23±1.48	98.35±0.34	95.61±0.62	95.61±0.91	35.73	3.72
+ EMCAM	64.75±5.03	65.21±3.41	66.16±3.05	81.90±3.44	98.60±0.22	98.60±0.16	95.90±1.63	98.33±1.31	12.72	1.35
+ ISA	65.27±4.75	66.10±3.14	65.90±3.71	82.22±4.64	97.10±0.66	97.25±0.28	94.80±1.64	94.80±1.20	25.60	2.87
+ CA	63.31±4.67	62.82±4.12	63.24±3.89	79.21±5.52	95.58±2.57	95.58±1.29	94.27±2.60	94.65±2.08	**9.84**	**0.91**
+ HSA	62.27±5.24	62.27±3.39	62.58±3.59	76.24±4.90	94.52±1.85	94.52±0.23	93.75±1.65	93.75±1.34	10.37	0.98
+ HySAM	65.73±3.46	66.85±3.55	67.08±2.76	82.57±3.77	98.60±1.01	98.60±0.93	95.90±1.07	97.95±0.65	10.57	1.25
+ HySAR (Ours)	**69.27**±2.53	**70.94**±1.51	**71.08**±2.49	**86.03**±3.17	**99.27**±0.43	**99.80**±0.41	**96.63**±0.78	**99.50**±0.46	13.73	1.32

the same plug-and-play replacement protocol to the state-space MAMBA-Tiny and convolutional Inception-V3 backbones, respectively. The consistent ordering across backbone families supports the backbone-agnostic character of HySAR and shows that its gains arise from hybrid-space shared representation refinement rather than backbone scaling.

ResNet18 backbone. Table 7 and Fig. 9 extend the comparison to additional ResNet18-based host multimodal fusion pipelines across six medical benchmarks. Under the same host-specific encoding blocks and prediction heads, HySAR consistently occupies the leading performance region across the evaluated multimodal fusion learners. This result shows that the observed improvements originate from hybrid-space shared representation refinement rather than redesigning the underlying host architecture.

MAMBA-Tiny and Inception-V3 backbones. Figures 10 and 11 evaluate the same replacement protocol with the state-space MAMBA-Tiny and convolutional Inception-V3 backbones, respectively. The favorable performance ordering is retained across the four survival cohorts and, for Inception-V3, across the medical-image classification benchmarks. These results demonstrate that the effectiveness of HySAR is not restricted to the inductive bias of a particular backbone family.

Medical classification across backbones. Figure 12 extends the comparison to ViT-Tiny- and Mamba-Tiny-based fusion on HAM10000 and SIPaKMeD. HySAR achieves the highest ACC and AUC among the compared refiners across all six host pipelines, supporting its effectiveness across transformer and state-space backbones.

Cross-backbone performance–efficiency trade-off. Figure 13 jointly compares average task performance with parameter count and GFLOPs for the MAMBA-Tiny and Inception-V3 backbones. HySAR-integrated learners remain in the high-performance, low-computation region across both backbone families. Because the modality encoders, host fusion pipeline, and prediction head are held fixed within each comparison, this shift isolates the favorable performance–efficiency behavior of the proposed shared representation learner.

Input-resolution sensitivity. Figure 14 evaluates MAIL+HySAR using 128×128, 224×224, 256×256, and 512×512 inputs while retaining the same backbone architecture and parameter count. The corresponding average C-index values are 75.18, 75.56, 75.42, and 76.03, re-

Table 8: Performance-based modality analysis of HySAR. CMC, MDG, FSG, and RRR denote reference-adjusted modality contribution, modality-dominance gap, fusion-synergy gain, and robustness-retention ratio, respectively.

	C-index				CMC (%)		MDG	FSG	RRR (%)		
Dataset	Omics	WSI	50% Both	All modalities	Omics	WSI	(% pts)	(% pts)	Omics	WSI	50% Both
BLCA	68.2	58.3	67.9	72.2	68.7	31.3	37.4	4.0	82.0	37.4	80.6
KIRP	81.5	65.9	76.7	85.8	66.5	33.5	32.9	4.3	88.0	44.4	74.6

spectively. Thus, increasing the input resolution improves average performance by only 0.24–0.86 points while increasing computation from 0.74 GFLOPs to 2.27–11.84 GFLOPs, corresponding to a 3.1×–16.0× increase. These results rule out brute-force spatial scaling as the principal source of improvement and identify 128 × 128 as the most favorable performance–efficiency operating point among the evaluated resolutions.

F.2 Performance-Based Modality Analysis

We quantify modality contribution, modality dominance, fusion synergy, and missing-modality robustness using predictive performance under single-, partial-, and all-modality inference. These measures characterize the predictive evidence retained by each modality and should not be interpreted as causal attribution or learned attention weights.

Let $S_{d,m}$ denote the C-index obtained on dataset d when only modality m is available, $S_{d,c}$ denote the score under availability condition c, $S_{d,\text{all}}$ denote the score when all modalities are available, and $L = 50$ denote the non-informative C-index reference.

Reference-adjusted modality contribution. We quantify the relative predictive contribution retained by modality m as

$$\text{CMC}_{d,m} = 100\frac{[S_{d,m} - L]_+}{\sum_{r\in\mathcal{M}_d}[S_{d,r} - L]_+}, \qquad [z]_+ = \max(z, 0), \tag{15}$$

where $\mathcal{M}_d$ denotes the modality set of dataset d. A larger $\text{CMC}_{d,m}$ indicates that modality m retains a greater share of the single-modality predictive evidence.

Modality-dominance gap. We measure the imbalance between the strongest and weakest modalities as

$$\text{MDG}_d = \max_{m\in\mathcal{M}_d} \text{CMC}_{d,m} - \min_{m\in\mathcal{M}_d} \text{CMC}_{d,m}. \tag{16}$$

A larger MDG_d indicates greater dependence on one modality relative to the other available modalities.

Fusion-synergy gain. We quantify the improvement obtained by using all modalities beyond the strongest individual modality as

$$\text{FSG}_d = S_{d,\text{all}} - \max_{m\in\mathcal{M}_d} S_{d,m}. \tag{17}$$

A positive FSG_d shows that multimodal fusion exploits complementary evidence beyond that retained by the strongest modality alone.

Robustness-retention ratio. For an availability condition c, we define

$$\text{RRR}_{d,c} = 100\frac{S_{d,c} - L}{S_{d,\text{all}} - L}. \tag{18}$$

This measure quantifies the percentage of reference-adjusted all-modality performance retained when modality availability is reduced.

Modality contribution and dominance. Table 8 shows that Omics retains the larger share of single-modality predictive evidence, contributing 68.7% on BLCA and 66.5% on KIRP, compared with 31.3% and 33.5% for WSI, respectively. The corresponding modality-dominance gaps of 37.4 and 32.9 points indicate a stronger single-modality imbalance on BLCA.

Table 9: Component and computational attribution under the fixed MAIL–ResNet18 setting. Δ performance, Δ#P, and the #F reduction are measured relative to native MAIL.

Configuration	Avg. C-index	Δ performance	#P (M)	Δ#P (M)	#F (G)	#F reduction (%)
Native MAIL	69.32	0.00	11.70	0.00	1.84	0.00
MILA only	72.13	+2.81	14.27	+2.57	0.74	59.78
DUHA only	71.38	+2.06	10.16	-1.54	0.65	64.67
Complete HySAR	75.18	+5.86	14.46	+2.76	0.74	59.78

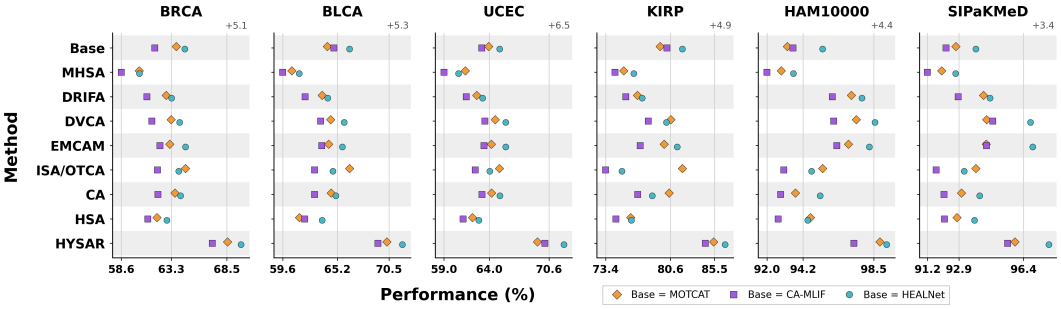

Figure 9: **Plug-and-play refiner comparison with ResNet18.** We compare refiners in MOTCAT (Xu & Chen, 2023), CA–MLIF (An et al., 2025), and HEALNet (Hemker et al., 2024) multimodal fusion pipelines across six medical benchmarks. Rows denote native configurations (i.e., Base) or replacement refiners; marker shapes identify host pipelines, and rightward positions indicate higher performance.

Fusion synergy. Despite the stronger Omics contribution, inference with all modalities improves C-index by 4.0 points on BLCA and 4.3 points on KIRP beyond their strongest individual modalities. These positive fusion-synergy gains demonstrate that HySAR exploits complementary multimodal evidence rather than merely relying on the dominant modality.

Robustness retention. Under Omics-only inference, HySAR retains 82.0% and 88.0% of the reference-adjusted all-modality performance on BLCA and KIRP, respectively. The corresponding WSI-only retention values are 37.4% and 44.4%, while the 50% Both setting retains 80.6% and 74.6%. These results characterize how predictive evidence is preserved under different availability conditions and complement the raw missing-modality C-index comparison.

F.3 Component and Computational Attribution

We further examine the contribution and computational behavior of the proposed components under the fixed MAIL–ResNet18 configuration. Average C-index is computed across BRCA, BLCA, UCEC, and KIRP; parameter and FLOP differences are measured relative to native MAIL.

Table 10: Descriptor-statistics (i.e., GMP, GSP, GMN, and GAP) ablation in DUHA.

Descriptor	BRCA	BLCA	UCEC	KIRP
GAP only	69.17	69.75	70.51	84.77
GAP + GMP	69.53	70.38	70.51	85.10
GAP + GMP + GSP	70.14	71.51	71.63	85.37
GAP + GMP + GSP + GMN	**70.52**	**72.18**	**72.21**	**85.79**

Component complementarity. Table 9 shows that MILA and DUHA independently improve average C-index by 2.81 and 2.06 points, respectively, whereas their integration through HFuse improves it by 5.86 points. The complete gain exceeds the sum of the two isolated gains by 0.99 points, demonstrating that cross-geometric interaction provides complementary benefit beyond independent component aggregation.

Computational behavior. The complete HySAR configuration requires 14.46M parameters and 0.74 GFLOPs, reducing computation by 59.78% relative to native MAIL. The DUHA-only configuration is the lightest variant, requiring 10.16M parameters and 0.65 GFLOPs. Completing HySAR from the MILA-only configuration adds 0.19M parameters without changing the reported GFLOPs at the displayed precision. Thus, the full hybrid-space design obtains its strongest performance with limited additional cost over its Euclidean branch.

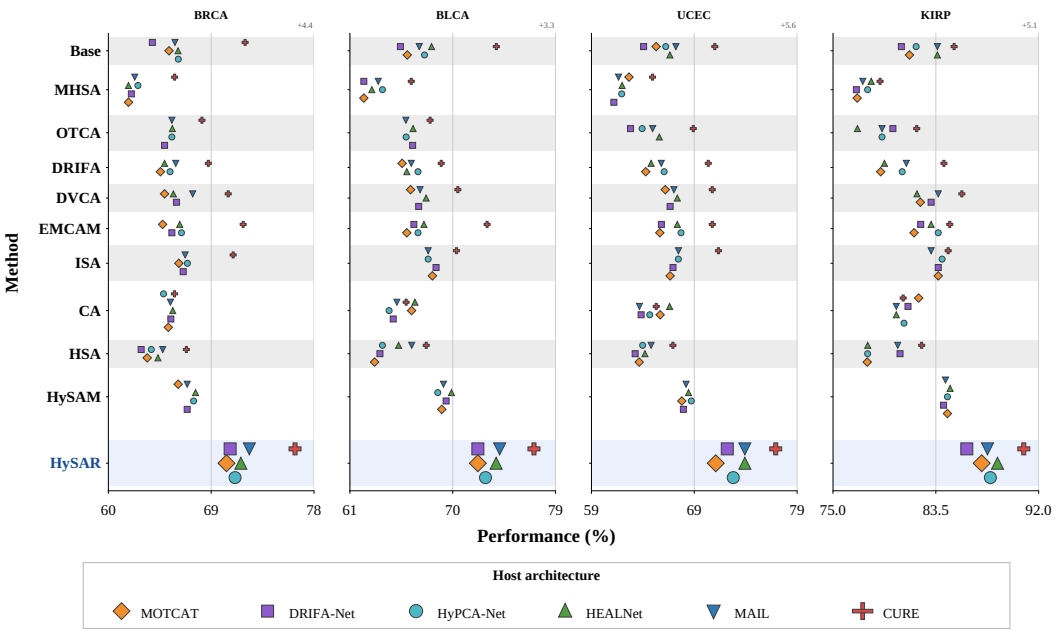

Figure 10: **Plug-and-play refiner comparison with `Mamba-Tiny`.** Panels show `C-index` on `BRCA`, `BLCA`, `UCEC`, and `KIRP` across six host multimodal fusion pipelines. Rows denote native configurations (`Base`) or replacement refiners; marker shapes identify host pipelines. Rightward positions indicate higher performance.

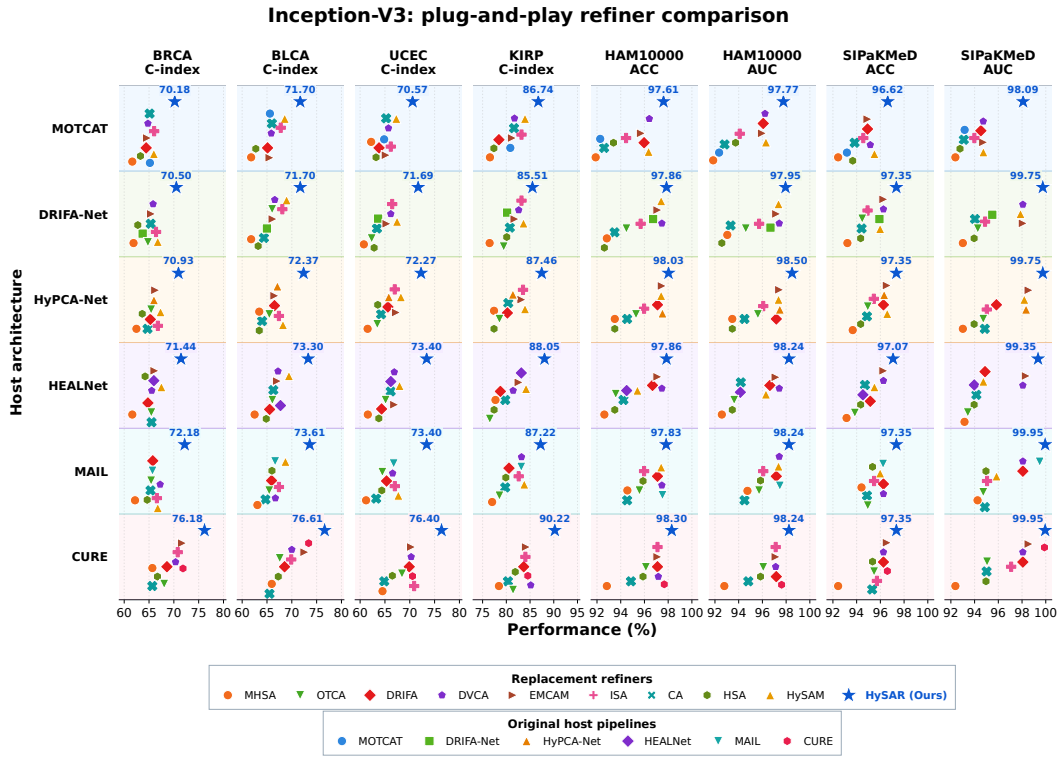

Figure 11: **Plug-and-play refiner comparison with `Inception-V3`.** Panels show `C-index` on `BRCA`, `BLCA`, `UCEC`, and `KIRP`, and `ACC`/`AUC` on `HAM10000` and `SIPaKMeD`. Rows identify host pipelines; marker colors and shapes identify refiners, with stars denoting `HySAR`. Rightward positions indicate higher performance.

F.4 Descriptor-Statistics Diversity

Descriptor diversity. Table 10 shows that combining `GAP`, `GMP`, `GSP`, and `GMN` improves average `C-index` by 1.63 points over `GAP` alone. `GAP` captures average contextual evidence, `GMP` empha-

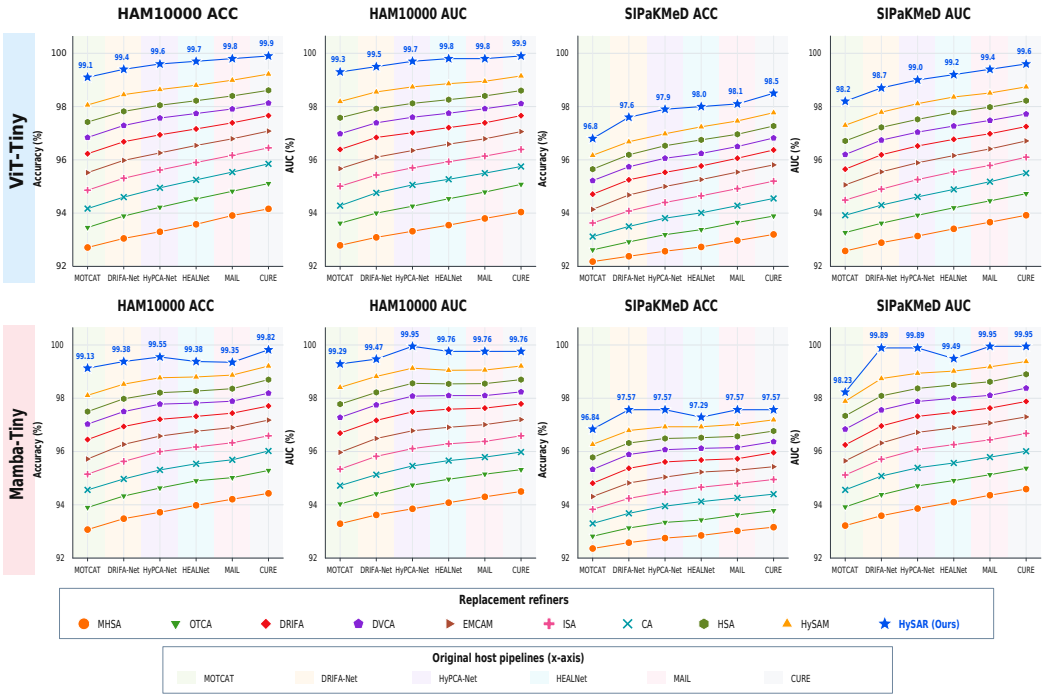

Figure 12: **Plug-and-play cross-modal refiner evaluation under `ViT-Tiny`- and `Mamba-Tiny`-based multimodal fusion pipelines.** For each of the `MOTCAT`, `DRIFA-Net`, `HyPCA-Net`, `HEALNet`, `MAIL`, and `CURE` host pipelines, we evaluate its native refiner and the applicable Euclidean-space, non-Euclidean-manifold, and proposed hybrid-space replacement refiners. The panels report `ACC` and `AUC` on `HAM10000` and `SIPaKMeD`. Curves and marker shapes distinguish replacement refiners, while the x-axis denotes host pipelines; stars denote `HySAR`.

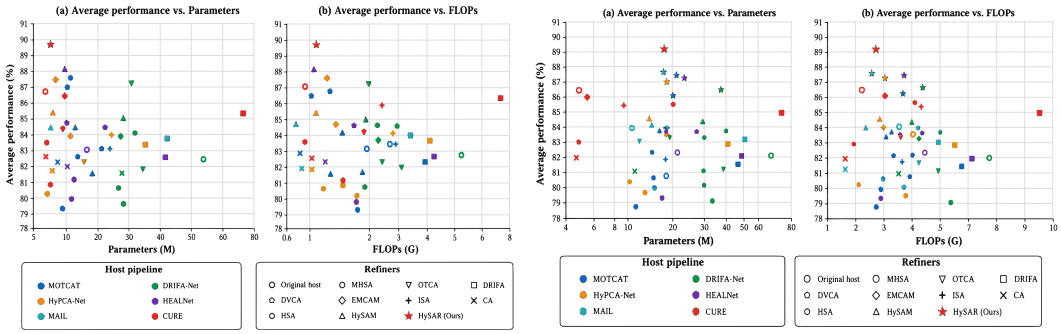

Figure 13: **Performance–efficiency trade-offs across backbones.** The left and right groups use `Mamba-Tiny` and `Inception-V3`, respectively. Within each group, panels **(a)** and **(b)** plot overall performance against parameters and `GFLOPs`, respectively. Overall performance averages the applicable `C-index`, `ACC`, and `AUC` scores across the medical benchmarks. Colors, marker shapes, and stars denote host pipelines, refiners, and `HySAR`; upper-left positions indicate a more favorable trade-off.

sizes salient activation peaks, `GSP` preserves total response magnitude, and `GMN` contributes complementary low-response structure. Their integration provides a more complete channel descriptor, reduces dependence on a single activation regime, and strengthens geometry-wise calibration without modifying the backbone.

F.5 Dataset-Relative Performance Analysis

Raw scores across heterogeneous benchmarks are influenced by both the performance already achieved by existing methods and the additional benefit provided by `HySAR`. We therefore report a dataset-relative analysis that separates these two factors. This analysis contextualizes performance

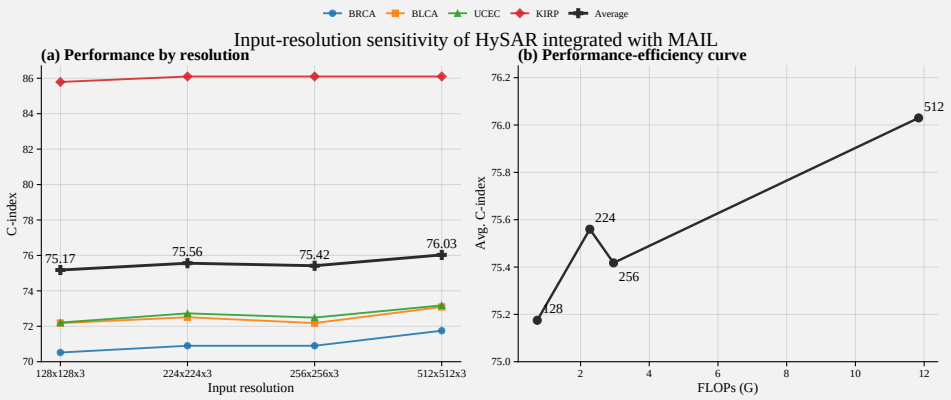

Figure 14: **Input-resolution sensitivity of HySAR integrated with MAIL.** Within MAIL, **(a)** reports dataset-wise and average C-index at 128×128, 224×224, 256×256, and 512×512 input resolutions, while **(b)** plots the average C-index against GFLOPs. The comparison characterizes the performance–efficiency trade-off as spatial resolution increases.

within each dataset and metric; it is not intended to rank the intrinsic difficulty of fundamentally different tasks or evaluation measures.

For dataset d and metric t, let $S^H_{d,t}$ denote the strongest reported HySAR result and

$$B_{d,t} = \max_{M \in \mathcal{B}} S_{d,t}(M) \tag{19}$$

denote the strongest non-HySAR result among the compared configurations $\mathcal{B}$.

Performance gain. The additional performance provided by HySAR is

$$\mathrm{AG}_{d,t} = S^H_{d,t} - B_{d,t}. \tag{20}$$

Reference headroom. The metric headroom remaining above the strongest compared baseline is

$$\mathrm{RH}_{d,t} = 100 - B_{d,t}. \tag{21}$$

Remaining-headroom closure. We quantify the fraction of the remaining metric range closed by HySAR as

$$\mathrm{RHC}_{d,t} = 100 \frac{\mathrm{AG}_{d,t}}{\mathrm{RH}_{d,t}}. \tag{22}$$

RHC should be interpreted jointly with absolute gain because near-ceiling benchmarks contain substantially less remaining headroom.

Reference performance and score differences. Table 11 shows that the strongest prior configurations on HAM10000, SIPaKMeD, and UPMC Food-101 already attain scores between 93.80 and 99.67, leaving only 0.33–6.20 points of reference headroom. Their high raw HySAR scores therefore partly reflect benchmark regimes in which the compared methods already operate near the corresponding metric ceiling.

In contrast, the four survival cohorts retain 16.15–30.20 points of reference headroom, while NYUDv2 and SUN-RGB-D retain 38.80 and 44.10 points, respectively. Their lower raw scores therefore coincide with lower performance across the compared methods rather than a smaller contribution from HySAR.

HySAR gain across benchmark regimes. Despite the larger reference headroom of the survival and RGB-D tasks, HySAR improves C-index by 3.04–5.41 points across the four survival cohorts and improves mIoU by 4.50 and 5.40 points on NYUDv2 and SUN-RGB-D, respectively. By comparison, the smaller absolute improvements on several classification benchmarks occur when the strongest prior configurations already operate close to the metric ceiling.

Near-ceiling performance. Absolute gain and remaining-headroom closure provide complementary interpretations. For example, the 0.79-point improvement on HAM10000 closes 63.20% of

the 1.25 points remaining above the strongest non-HySAR configuration. Similarly, the 2.70-point improvement on UPMC Food−101 closes 43.55% of its remaining reference headroom. Thus, a smaller absolute gain on a near-ceiling benchmark does not necessarily indicate a weaker representation-learning effect.

Difference decomposition for common metrics. For two datasets d and k evaluated using the same metric t, their HySAR score difference can be decomposed as

$$\underbrace{S^H_{d,t} - S^H_{k,t}}_{\text{HySAR score difference}} = \underbrace{B_{d,t} - B_{k,t}}_{\text{reference-performance difference}} + \underbrace{\mathrm{AG}_{d,t} - \mathrm{AG}_{k,t}}_{\text{HySAR-gain difference}}. \tag{23}$$

This decomposition identifies whether a score gap is inherited from the performance level of the compared methods or arises from a difference in the additional benefit provided by HySAR.

For example, the C−index difference between KIRP and UCEC is

$$88.79 - 75.21 = (83.85 - 69.80) + (4.94 - 5.41) = 13.58.$$

The strongest-reference difference is 14.05 points, whereas HySAR provides 0.47 points more absolute gain on UCEC. Thus, the lower raw score on UCEC is inherited from the lower performance level of all compared configurations, not from a weaker incremental contribution of HySAR.

Similarly, the mIoU difference between NYUDv2 and SUN−RGB−D is

$$65.70 - 61.30 = (61.20 - 55.90) + (4.50 - 5.40) = 4.40.$$

Although SUN−RGB−D has the lower raw score, HySAR provides 0.90 points more absolute gain on SUN−RGB−D. This demonstrates why raw cross-dataset scores should be interpreted jointly with the corresponding reference performance and method-specific gain.

G Discussion

G.1 Performance–Efficiency–Robustness Balance

Controlled attribution to shared representation refinement. Across all host–refiner comparisons, only the cross-modal shared representation refiner is replaced, while the backbone, modality-specific encoding blocks, remaining host fusion pipeline, prediction head, data partition, and training protocol remain fixed. The consistent improvements therefore isolate the contribution of hybrid-space shared representation refinement rather than architectural scaling. Across ten medical and natural-domain multimodal benchmarks, HySAR-integrated learners improve performance by 4.8 points on average, reduce parameters and FLOPs by up to 84.1% and 88.0%, respectively, relative to cost-intensive alternatives, and improve missing-modality C−index by up to 8.7 points. Within the evaluated host–refiner configurations, these results establish the strongest observed joint performance–efficiency–missing-modality-robustness balance.

Comparison with the strongest prior replacement refiner. Under the fixed MAIL–ResNet18 survival setting, HySAM is the strongest prior replacement refiner, obtaining an average C−index of 70.38 with 11.15M parameters and 0.62 GFLOPs. Complete HySAR reaches 75.18, yielding an exact 4.80-point improvement for an additional 3.31M parameters and only 0.12 GFLOPs. Hence, HySAR does not obtain its performance through substantial computational expansion; rather, it exchanges a modest increase over the lowest-cost alternative for a markedly stronger shared representation.

Why efficiency does not sacrifice representation quality. HySAR allocates computation to structured operations instead of dense high-dimensional attention matrices. MILA employs depth-wise, group-wise, and dilated depth-wise homogeneous convolutions to preserve multi-scale Euclidean evidence. DUHA applies its Poincaré and Proper-Velocity transformations after global channel summarization and low-rank projection, avoiding manifold processing over every pair of spatial tokens. HFuse subsequently operates on channel-wise geometric responses rather than constructing a dense spatial correspondence matrix. This organization strengthens representational expressiveness while retaining a compact computational profile.

G.2 MECHANISTIC INTERPRETATION OF HySAR

Complementarity of Euclidean and dual-hyperbolic evidence. The performance gain arises from coordinated geometric refinement rather than from one isolated component. MILA preserves channel-local, group-structured, and context-expanded Euclidean evidence through MDHC, MGHC, and MDDHC, respectively. DUHA complements these cues by learning curvature-aware hierarchical responses through the bounded Poincaré-ball formulation in PARL and the unconstrained Proper-Velocity formulation in PVRL. HFuse then performs adaptive cross-geometric exchange across modalities instead of treating these responses as independent predictions.

Table 11: Dataset-relative performance analysis of Tables 1 and 2. For each dataset–metric pair, S^H and B denote the strongest HySAR and non-HySAR scores, respectively; AG, RH, and RHC denote absolute gain, reference headroom, and remaining-headroom closure.

Dataset	Metric	S^H	B	AG	RH	RHC (%)
BRCA	C-index	74.52	70.39	4.13	29.61	13.95
BLCA	C-index	75.18	72.14	3.04	27.86	10.91
UCEC	C-index	75.21	69.80	5.41	30.20	17.91
KIRP	C-index	88.79	83.85	4.94	16.15	30.59
HAM10000	ACC	99.54	98.75	0.79	1.25	63.20
HAM10000	AUC	99.54	98.75	0.79	1.25	63.20
SIPaKMeD	ACC	96.63	95.97	0.66	4.03	16.38
SIPaKMeD	AUC	99.81	99.67	0.14	0.33	42.42
NYUDv2	mIoU	65.70	61.20	4.50	38.80	11.60
SUN-RGB-D	mIoU	61.30	55.90	5.40	44.10	12.24
MM-IMDb	F1-score	76.40	74.10	2.30	25.90	8.88
UPMC Food-101	ACC	96.50	93.80	2.70	6.20	43.55

Under the fixed MAIL–ResNet18 configuration, MILA-only and DUHA-only improve C-index by 2.81 and 2.06 points on average, respectively, whereas complete HySAR improves it by 5.86 points. The complete gain exceeds the sum of the isolated gains by 0.99 points, supporting non-additive cross-geometric complementarity. Moreover, jointly using MDHC, MGHC, and MDDHC improves C-index by 3.80 points on average, confirming that the streams capture complementary rather than redundant Euclidean evidence.

Why parallel geometry learning is preferable. Parallel processing allows PARL and PVRL to operate on the same unfiltered contextual descriptor. A cascaded implementation instead requires the second hyperbolic formulation to process a representation already transformed by the first, potentially attenuating formulation-specific evidence before fusion. Parallel DUHA improves over the individual hyperbolic formulations by 0.87–1.46 points, while parallel DUHA and parallel HySAR outperform their cascaded counterparts by 0.58 and 0.64 points on average, respectively, at identical measured parameter and FLOP counts. The improvement therefore reflects better information preservation rather than additional capacity.

Role of diverse channel statistics. The global average, maximum, sum, and minimum statistics used by DUHA characterize complementary response regimes. GAP captures average contextual evidence, GMP emphasizes salient activation peaks, GSP preserves total response magnitude, and GMN contributes low-response complementary structure. Their complete integration improves average C-index by 1.63 points over GAP alone, indicating that geometry-aware calibration benefits from a descriptor that does not depend on one activation statistic.

G.3 MISSING-MODALITY RELIABILITY AND MODALITY SYNERGY

Availability-aware geometric refinement. Missing-modality reliability follows from the interaction of modality-local and cross-modal mechanisms. MILA and DUHA first preserve Euclidean and hierarchy-aware evidence within each observed modality. HFuse then masks unavailable responses before softmax normalization and renormalizes the remaining valid weights. Consequently, a missing branch neither absorbs probability mass nor perturbs the representations learned from the observed modalities. Across Omics-only, WSI-only, 50% Both, and all-modality inference, HySAR improves C-index by 5.2–8.7 points on BLCA and by up to 6.3 points on KIRP.

Fusion complements rather than selects the dominant modality. Omics retains the larger share of single-modality predictive evidence on both BLCA and KIRP. Nevertheless, all-modality inference improves C-index by 4.0 and 4.3 points beyond the strongest individual modality on these cohorts, respectively. These positive fusion-synergy gains show that HySAR does not merely select the dominant input; it extracts additional predictive evidence from the weaker but complementary modality. Under Omics-only inference, it retains 82.0% and 88.0% of the reference-adjusted all-modality performance on BLCA and KIRP, respectively.

Table 12: **Leading-order asymptotic complexity of HySAR and representative cross-modal refiners. (a)** Component-wise analysis of HySAR; **(b)** comparison with representative prior refiners. The common backbone, modality encoders, and prediction head are excluded.

(a) HySAR component-wise complexity

Module	Main computation	Time complexity	Parameter and working-memory complexity
MILA	Multi-scale depth-/group-wise convolutions and low-rank channel projection	$\mathcal{O}(PN\Gamma_C + (P+1)Cd)$	#P : $\mathcal{O}(P\Gamma_C + (P+1)Cd)$ Mem. : $\mathcal{O}(NC)$
DUHA	Global channel descriptors, low-rank projections, PARL, and PVRL	$\mathcal{O}\big(NC + Cd + C^2\big)$	#P : $\mathcal{O}(Cd + C^2)$ Mem. : $\mathcal{O}(C + d)$
HFuse	Availability-aware cross-geometric channel gating	$\mathcal{O}\big(M^2C + MNC\big)$	#P : $\mathcal{O}(MC)$ Mem. : $\mathcal{O}(M^2C)$
HySAR	Complete Euclidean–dual-hyperbolic refinement and fusion	$\mathcal{O}\big(MPN\Gamma_C + MNC + M(P+2)Cd + MC^2 + M$	#P : $\mathcal{O}\big(MP\Gamma_C + M(P+2)Cd + MC^2 + MC\big)$ Mem. : $\mathcal{O}(MNC + M^2C)$

(b) Representative prior cross-modal refiners

Refiner	Main computation	Time complexity	Parameter and working-memory complexity
MHSA	Dense attention over $S = \sum_i N_i$ multimodal tokens	$\mathcal{O}\big(SC^2 + S^2C\big)$	#P : $\mathcal{O}(C^2)$ Mem. : $\mathcal{O}(S^2)$
OTCA	Dense transport matrix with U Sinkhorn updates	$\mathcal{O}(N_hN_g(C+U))$	#P : $\mathcal{O}(C^2)$ Mem. : $\mathcal{O}(N_hN_g)$
ISA	R iterative cross-attention rounds using L shared latent tokens	$\mathcal{O}\big(RM\big[NLC + (N+L)C^2\big]\big)$	#P : $\mathcal{O}(MC^2)$ Mem. : $\mathcal{O}(MNL)$
HySAM	Fast DCT, Poincaré–Lorentz processing, quantum mixing, and gating	$\mathcal{O}(Q[NC\log N + e])$	#P : $\mathcal{O}(Qe)$ Mem. : $\mathcal{O}(NC + e)$

Notation. M denotes the number of available modalities; $N = H_jW_j$ the number of spatial positions; C the channel dimension; P the number of MILA stages; d the low-rank dimension; and Γ_C the grouped multi-scale convolution cost defined in Eq. 24. Furthermore, $S = \sum_i N_i$; N_h and N_g denote histology and genomic token counts; U the number of Sinkhorn updates; L and R the number of shared latent tokens and attention rounds; and e and Q the HySAM descriptor dimension and number of calls, respectively. All expressions denote leading-order operator complexity.

G.4 Generalization Across Architectures and Tasks

Cross-host and cross-backbone consistency. The favorable performance ordering persists across multiple host multimodal fusion architectures and four backbone settings: ResNet18, Inception-V3, ViT-Tiny, and MAMBA-Tiny, spanning convolutional, transformer, and state-space families. Because the host pipeline is fixed within each replacement comparison, this consistency indicates that the gains originate from HySAR's shared representation refinement rather than the inductive bias or capacity of one backbone family.

Cross-domain and cross-task transfer. HySAR improves WSI–omics survival prediction, image–tabular medical classification, RGB-D semantic segmentation, and image–text classification. These settings differ in modality structure, output granularity, prediction objective, and evaluation metric. The consistent gains therefore support HySAR as a general cross-modal refiner rather than a task-specific modification designed only for medical survival analysis.

Spatial scaling is not the primary source of improvement. Increasing the input resolution from 128×128 to 512×512 improves average C-index from 75.18 to 76.03, corresponding to only 0.86 points, while increasing computation from 0.74 to 11.84 GFLOPs, i.e., by $16.0\times$. Accordingly, the principal performance gain arises from the organization of Euclidean, dual-hyperbolic, and cross-modal evidence rather than brute-force spatial scaling.

G.5 Asymptotic Complexity and Scalability

Notation and analysis scope. We analyze one HySAR block and exclude the common backbone, modality encoders, and prediction head. Let $M = |\mathcal{O}|$ denote the number of available modalities, $N = H_j W_j$ the number of spatial positions, and $C = C_j$ the channel dimension at block j. Let $P = 2$ denote the number of MILA stages, $d \ll C$ the low-rank bottleneck dimension, G the number of groups in GWC, and G_p the number of groups in GPC. For the kernel set $\mathcal{K} = \{1, 3, 5\}$, define

$$\kappa = \sum_{k\in\mathcal{K}} k^2 = 35, \qquad \Gamma_C = 2\kappa C + \frac{\kappa C^2}{G} + \frac{9C^2}{G_p}. \tag{24}$$

Here, $2\kappa C$ accounts for the multi-kernel depth-wise operations in MDHC and MDDHC, $\kappa C^2/G$ accounts for MGHC, and $9C^2/G_p$ accounts for the three grouped point-wise consolidations, each mapping concatenated $3C$ responses to C channels. The analysis gives leading-order operator counts; constants, normalization, activations, residual additions, and channel shuffling are absorbed into the corresponding linear terms. Table 12(a) summarizes the time, parameter, and working-memory complexity of HySAR, while Table 12(b) compares representative refiners. For fixed channel width and modality count, HySAR scales linearly with spatial size without constructing a quadratic spatial attention matrix.

MILA complexity. For one modality and one stage, MDHC and MDDHC require $\mathcal{O}(N\kappa C)$ each, while MGHC requires $\mathcal{O}(N\kappa C^2/G)$. The three GPC consolidations require $\mathcal{O}(9NC^2/G_p)$. Stage-wise pooling and IMF require $\mathcal{O}(NC)$, while a factorized projection $PC \to d \to C$ requires $\mathcal{O}((P+1)Cd)$. Therefore,

$$\begin{aligned} \mathcal{T}_{\text{MILA}} &= \mathcal{O}(PN\Gamma_C + (P+1)Cd)\,, \\ \mathcal{P}_{\text{MILA}} &= \mathcal{O}(P\Gamma_C + (P+1)Cd)\,, \\ \mathcal{S}_{\text{MILA}} &= \mathcal{O}(NC + PC), \end{aligned} \tag{25}$$

where $\mathcal{T}$, $\mathcal{P}$, and $\mathcal{S}$ denote time, trainable parameter, and working-memory complexity, respectively. Thus, MILA is linear in spatial size N; its channel-order term is controlled by group-wise processing.

DUHA complexity. The four global pooling operators require $\mathcal{O}(NC)$ operations. The two low-rank projections require $\mathcal{O}(Cd)$, while PARL's norm, exponential mapping, and channel modulation require $\mathcal{O}(C)$. In the current formulation, PVRL contains an intrinsic $C \to C$ PVFC transformation, yielding a conservative $\mathcal{O}(C^2)$ term; its exponential and logarithmic maps remain linear in C. Geometry-wise softmax gating and sigmoid calibration also require $\mathcal{O}(C)$. Hence,

$$\begin{aligned} \mathcal{T}_{\text{DUHA}} &= \mathcal{O}(NC + Cd + C^2), \\ \mathcal{P}_{\text{DUHA}} &= \mathcal{O}(Cd + C^2), \\ \mathcal{S}_{\text{DUHA}} &= \mathcal{O}(C + d). \end{aligned} \tag{26}$$

When the intrinsic PVFC is implemented through a rank-d factorization, the C^2 term becomes Cd.

HFuse complexity. For each ordered target–source pair (i, r) with $i \neq r$, HFuse forms and gates a C-dimensional cross-geometric response. The $M(M-1)$ pairwise responses therefore require $\mathcal{O}(M^2C)$ operations. Broadcasting the resulting attention maps over the encoded representations requires $\mathcal{O}(MNC)$ operations. Thus,

$$\begin{aligned} \mathcal{T}_{\text{HFuse}} &= \mathcal{O}(M^2C + MNC), \\ \mathcal{P}_{\text{HFuse}} &= \mathcal{O}(MC), \\ \mathcal{S}_{\text{HFuse}} &= \mathcal{O}(M^2C). \end{aligned} \tag{27}$$

Pairwise responses can be accumulated sequentially, reducing the additional fusion memory from $\mathcal{O}(M^2C)$ to $\mathcal{O}(MC)$ without changing the arithmetic order.

Complete `HySAR` complexity. Combining the modality-wise `MILA` and `DUHA` operations with `HFuse` gives

$$\begin{aligned}\mathcal{T}_{\texttt{HySAR}} &= \mathcal{O}\left(MPN\Gamma_C + MNC + M(P+2)Cd + MC^2 + M^2C\right),\\ \mathcal{P}_{\texttt{HySAR}} &= \mathcal{O}\left(MP\Gamma_C + M(P+2)Cd + MC^2 + MC\right),\\ \mathcal{S}_{\texttt{HySAR}} &= \mathcal{O}\left(MNC + M^2C\right).\end{aligned} \tag{28}$$

For fixed P, kernel set, and modality count, `HySAR` is linear in the number of spatial positions and does not construct an $\mathcal{O}(N^2)$ spatial attention matrix. If G and G_p grow proportionally with C, the grouped convolution terms also approach linear channel scaling.

Interpretation relative to prior refiners. Dense token-pair refiners such as `MHSA` incur quadratic $\mathcal{O}(S^2)$ attention memory, while optimal-transport co-attention requires a dense cross-modal transport matrix before mini-batch approximation. `ISA` avoids full token–token self-attention by using a shared latent bottleneck, but its cost grows with the number of latent tokens and iterative updates. Both `HySAM` and `HySAR` avoid an N^2 spatial affinity matrix. However, their mechanisms differ: `HySAM` applies frequency-domain transformation followed by hyperbolic–quantum mixing, whereas `HySAR` explicitly preserves multi-scale Euclidean evidence and cross-fuses it with dual-hyperbolic responses.

The asymptotic comparison should therefore not be interpreted as claiming that `HySAR` is the lowest-cost refiner in every setting. Under fixed `MAIL-ResNet18`, `HySAM` is 0.12 `GFLOPs` cheaper, but `HySAR` improves average `C-index` by 4.80 points. The empirical advantage is thus a substantially stronger representation for modest additional cost, rather than a reduction in arithmetic order alone.

G.6 Societal Impact

Accessible multimodal `AI`. Large multimodal architectures can be difficult to train and deploy in hospitals, laboratories, and research groups with limited accelerators, memory, or energy budgets. By replacing only the shared representation refiner and retaining the validated host pipeline, `HySAR` can improve multimodal prediction without requiring an entirely new high-capacity architecture. Its favorable parameter and `FLOP` profile can therefore broaden access to multimodal learning in resource-constrained clinical, scientific, and educational environments.

More inclusive use of incomplete multimodal records. Complete-case analysis can exclude patients or samples lacking costly imaging, omics, depth, text, or tabular measurements. Availability-aware refinement allows observed evidence to remain useful when one modality is unavailable, potentially reducing the systematic exclusion of incompletely observed records. This property is particularly relevant when access to advanced molecular assays or imaging technologies varies across institutions and patient populations. Missing-modality robustness is intended to preserve prediction continuity when information is unavailable, rather than to replace clinically necessary measurements.

Computational sustainability and model reuse. The plug-and-play design encourages reuse of existing modality encoders, fusion pipelines, and task heads, reducing the engineering and computational burden associated with training separate end-to-end architectures for every dataset or task. Lower computation per forward pass can also reduce memory demand and energy use during repeated experimentation and deployment. Accordingly, `HySAR`'s societal value follows not only from predictive improvement, but also from enabling more resource-conscious and broadly deployable multimodal representation learning.