MORE THAN A SHORTCUT: A HYPERBOLIC APPROACH TO EARLY-EXIT NETWORKS

APPENDIX

A. INHERENT HIERARCHY IN PRE-TRAINED BACKBONES

Our proposed method, Hyperbolic Early-Exit networks (`HypEE`), reframes the multi-stage architecture by leveraging hyperbolic geometry to explicitly model the hierarchical relationships between intermediate network layers. Instead of treating early exits as a series of independent classifiers, we consider them a sequence of refinement stages operating within a geometrically structured latent space. Below, we first provide an empirical motivation for this geometric shift.

A.1. Revealing Hierarchy within Intermediate Representations

The premise of our work is that the representations learned by deep backbone networks are inherently hierarchical across their depth. We analyze the geometric structure of intermediate embeddings from a pre-trained BEATs [1] audio backbone. We adopt the concept of Gromov's δ-hyperbolicity [2], a formal measure that quantifies the "tree-likeness" of a metric space. A low, scale-invariant δ-hyperbolicity value, denoted $\delta_{rel} \in [0, 1]$ [1], indicates that the space is highly tree-like and thus well-suited for embedding in a hyperbolic geometry [3].

We conduct an experiment where we extract embeddings from the backbone at different depths: 25% through the network, 50% through, and at the final layer (100%). We then compute δ_{rel} both within the set of embeddings from a single layer (*i.e.* intra-layer) and between the sets of embeddings from different layers (*i.e.* inter-layer). The results, summarized in Table 1, reveal two key findings. First, the intra-layer embeddings at each depth exhibit low δ_{rel} values (0.23-0.30), confirming that the representations for different audio samples are already organized in a hierarchical fashion. More importantly, the inter-layer hyperbolicity is even more pronounced, with δ_{rel} values as low as 0.143 between the 50% and 100% layers.

This strong empirical evidence suggests that a natural hierarchical structure exists not just among audio samples (inline with [3]'s observation for image samples), but critically, across the depth of the audio backbone itself. The representations at deeper layers are structurally related to those at shallower layers in a tree-like manner. This finding motivates our core proposal: to replace the geometrically unstructured Euclidean space of traditional Early-Exit models with a hyperbolic latent space, which provides a natural inductive bias for learning and preserving these hierarchical relationships.

Table 1. Gromov's δ-hyperbolicity for intermediate embeddings from a pre-trained BEATs backbone. We compare both intra-layer (top) and inter-layer (bottom) configurations. The significantly lower δ_{rel} values for inter-layer comparisons indicate a strong hierarchical structure across the network's depth and strongly motivate the use of hyperbolic geometry to model the network's depth-wise progression.

X	**Y**	δ_{rel}	c
25%	25%	0.282	0.26
50%	50%	0.304	0.223
100%	100%	0.233	0.379
25%	50%	0.247	0.338
25%	100%	0.148	0.94
50%	100%	0.143	1.012

Table 2. Detailed Early-Exit Analysis for Global Norm Exit and Classwise Norm Exit Strategies

Exit Strategy	**Gate**	**Triggered %**	**Correct %**	**Incorrect %**
Global Norm Exit	EE_0	35.61	80.52	19.48
	EE_1	36.74	78.13	21.87
	Final	27.65	60.18	39.82
Classwise Norm Exit	EE_0	30.05	98.82	1.18
	EE_1	39.08	99.73	0.27
	Final	30.87	61.81	38.19

B. DETAILED EARLY-EXIT TRIGGER ANALYSIS FOR `HypEE`

We further detail our breakdown of proposed EE triggers in Table 2. It is evident that a geometric trigger is exceptionally precise at identifying samples it can classify correctly: of the samples exited at EE_0 and EE_1, over 98.8% and 99.7% are classified correctly, respectively. The model intelligently offloads the truly difficult samples (approx. 31% of the total) to the final, most capable exit. This demonstrates that our geometry-aware triggering mechanism successfully operationalizes the learned hierarchy, completing the `HypEE` framework and delivering a superior accuracy-efficiency trade-off.

C. ADDITIONAL VISUALIZATION FOR HYPERBOLIC LATENTS

C.1. UMAP Visualization of Exit Gate Embeddings

In addition to the t-SNE plots in the main paper, we use UMAP (Uniform Manifold Approximation and Projection) to visualize the learned embeddings, as shown in Fig. 1. The embeddings from the three exit gates are projected from the *Lorentz* hyperboloid onto its equivalent *Poincaré* disk representation. The visualization, colored

[1] $\delta_{rel} = \frac{2\delta}{\text{diameter}}$
Diameter: maximal pairwise distance. Any latent space is considered δ-hyperbolic if, for some value δ, every point located on the edge of a geodesic triangle is within a distance of δ from another edge.

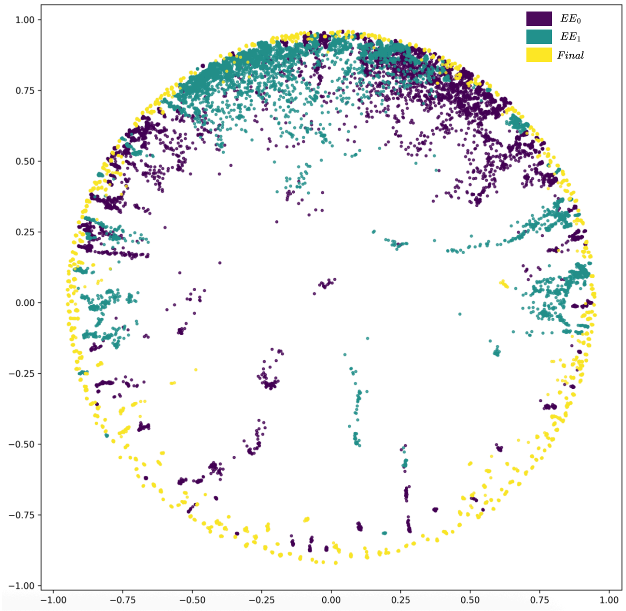

Fig. 1. UMAP [4] visualization of the learned hyperbolic embeddings from the SED model, projected onto the *Poincaré* disk. The embeddings are colored by their exit level (EE_0, EE_1, $Final$). The plot shows a clear radial hierarchy, with earlier exit embeddings positioned more centrally, providing evidence of the learned entailment structure.

by exit level, provides further evidence of the hierarchical structure imposed by our entailment loss. The embeddings from the first exit, EE_0 (purple), are predominantly located in the central region of the disk, representing higher uncertainty. The embeddings from the second exit, EE_1 (teal), extend outwards from this core, and the $Final$ exit embeddings (yellow) are pushed furthest towards the periphery. This clear radial separation confirms that the model learns a structured progression from general to specific representations across its depth.

C.2. Contextual Clustering with Hyperbolic k-means

To investigate the semantic organization of the learned space at the Early-Exits, we perform an unsupervised clustering experiment. Our hypothesis is that the Early-Exits learn to group sounds into broader, contextually relevant acoustic categories, even without explicit supervision to do so.

Specifically, we select five distinct, high-level acoustic concepts from the Audioset Strong evaluation set: Respiratory, Ringing, Speech, Singing, and Mechanical (engines), each comprising several fine-grained classes. We gather evaluation samples belonging to these classes and apply hyperbolic k-means clustering (k=5) to their embeddings from EE_0 and EE_1 separately.

Fig. 2 shows the proportion of each fine-grained class within the emergent clusters found by k-means. The results reveal a remarkable correspondence between the unsupervised clusters and our predefined semantic groups. For example, at EE_0, Cluster 0 is overwhelmingly composed of various speech and singing classes (human vocalizations), while Cluster 3 is almost exclusively made up of different types of bell and chime sounds (high-frequency alerts/musical sounds). Similarly, a significant portion of engine-related sounds is grouped into Cluster 2.

Implication for Contextual AI. This emergent clustering demonstrates that the Early-Exits in `HypEE` learn a meaningful acoustic taxonomy. EE_0 can effectively distinguish between high-level concepts like “human vocalizations” or “mechanical noise” even if it remains uncertain about the specific subclass. This capability is highly valuable for contextual AI on resource-constrained devices. An “always-on” system could use a computationally cheap Early-Exit to make a broad contextual inference (e.g., “human presence detected,” “vehicle nearby”) and only trigger the more expensive, deeper layers when a fine-grained classification is required, enabling a more intelligent and efficient allocation of resources [5].

D. LOOKAHEAD PREDICTION VIA ENTAILMENT CONES

In Algorithm 1 (main paper), we demonstrated a triggering mechanism based on the norm of hyperbolic embeddings, which serves as a proxy for uncertainty. Beyond this, we explore whether the entailment cone itself—the core of our hierarchical training objective—could be directly harnessed for inference. Inspired by work on predicting uncertain futures [6], where hyperbolic models “hedge their bets” by forecasting a more abstract outcome, we investigate if an embedding at an Early-Exit, h_i, could “forecast” its final classification by examining the classes of more refined embeddings that are geometrically consistent with it.

Specifically, we define “geometric consistency” as falling within the entailment cone. We designed an experiment where each sample from the ESC-50 validation set acts as a “query” represented by its embedding at the first exit, EE_0. A “reference set” consists of all training set embeddings from the subsequent, more refined exits (EE_1 and $Final$). For each query, we identify all reference embeddings that fall within its entailment cone, a process conceptually illustrated in Fig. 3. Since the entailment loss is non-zero during our training, we relax the strict condition with a threshold T, such that a reference sample h_{ref} is considered to be within the cone of a query h_{query} if $ext(h_{query}, h_{ref}) \leq T \cdot aper(h_{query})$.

Results and Future Directions. Fig. 4 shows that at tight thresholds (e.g., $T = 1.2$), the precision is remarkably high: 93.2% of the reference samples retrieved from EE_1 share the same ground-truth class as the query sample. This indicates that the entailment cone is semantically coherent and contains strong predictive information about the query’s identity. As the threshold is relaxed, the number of retrieved samples increases, but precision naturally decreases.

While promising, we present this as an exploratory analysis rather than a practical inference algorithm due to two main challenges. First, the computational cost of comparing a query against a large reference set is prohibitive for real-time applications. Second, many query samples do not retrieve any reference samples at stricter thresholds, limiting the coverage of the method. However, this exploration successfully validates the rich, predictive structure of the `HypEE` latent space and opens several exciting avenues for future work. A key direction would be to develop methods to make this *look ahead* approach practical, perhaps by learning a small, representative set of prototype reference embeddings to reduce the search space, or by training a model to directly predict the class distribution within an embedding’s entailment cone. Our initial result strongly suggests that the geometry learned by `HypEE` is not just a representational artifact, but a potentially powerful tool for future inference strategies.

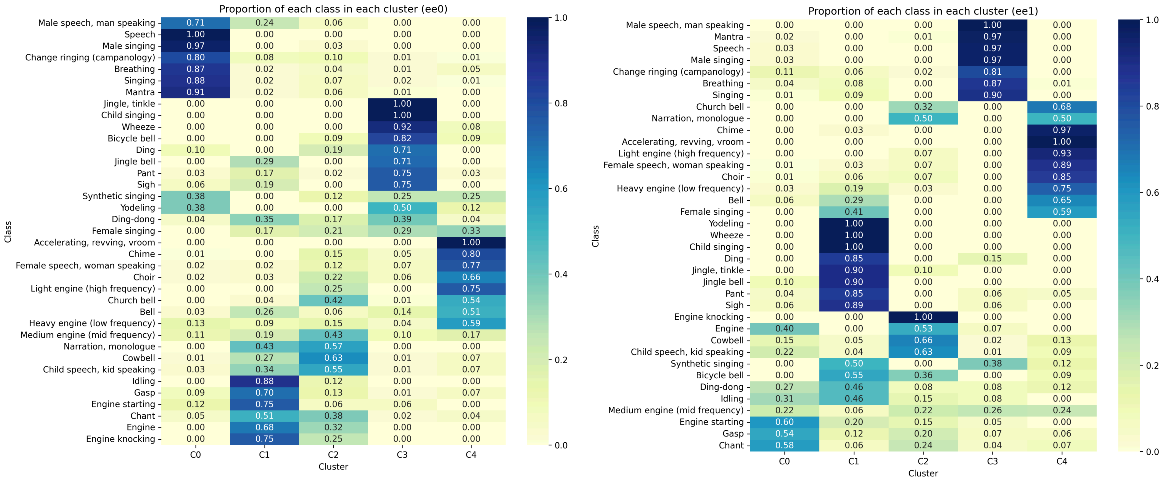

Fig. 2. Proportion of hand-picked Audioset Strong classes within each of the 5 clusters discovered by hyperbolic k-means, for embeddings from EE_0 (left) and EE_1 (right). The unsupervised clusters show a strong correspondence with high-level acoustic concepts (e.g., human speech, bells, engines), indicating that the Early-Exits learn a meaningful contextual hierarchy.

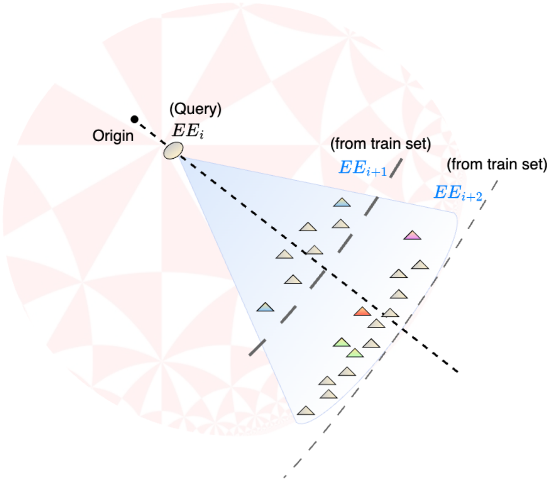

Fig. 3. A conceptual illustration of the *look ahead* prediction strategy. A query sample's embedding at an Early-Exit, EE_i, defines an entailment cone. We *look ahead* by identifying reference embeddings from the training set of subsequent exits (e.g., EE_{i+1}, $Final$) that fall within this cone. The ground-truth classes of these retrieved reference samples are then used to forecast the query's most likely class.

E. QUALITATIVE ANALYSIS OF THE LEARNED HIERARCHY VIA TRAVERSAL

To qualitatively evaluate the hierarchical structure learned by HypEE, we conduct a traversal experiment inspired by recent work in hyperbolic representation learning [7, 8]. The objective is to analyze the path from a specific, fine-grained embedding (from the final exit) to the most general concept in the latent space (the '[ROOT]'). A well-structured hierarchy should reveal a smooth progression from specific to abstract concepts along this path.

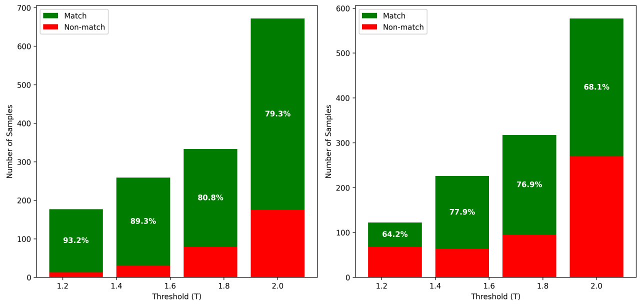

Fig. 4. Results of the *look ahead* prediction experiment. For different entailment cone thresholds (T), we show the number of retrieved reference samples from later exits that either match (green) or do not match (red) the query sample's ground-truth class. The percentages indicate the precision (match / total retrieved). The left and right plots correspond to using reference samples from EE_1 and the $Final$ exit, respectively.

Beyond providing qualitative validation, these findings point towards several practical applications for the structured latent space learned by HypEE. The navigable hierarchy offers a powerful tool for model interpretability and error analysis, allowing researchers to trace the refinement process for a given input. Furthermore, the emergent acoustic taxonomy at the earliest exit could enable more sophisticated, context-aware triggering mechanisms. For instance, an "always-on" device could use the computationally cheap EE_0 to make broad contextual inferences (e.g., detecting a "transient event") and only activate the deeper, more power-intensive exits when a fine-grained classification is necessary. This opens avenues for designing more efficient and intelligent sensing systems that leverage a deeper understanding of their acoustic environment.

F. REFERENCES

[1] Chen Sanyuan et al., “BEATs: Audio pre-training with acoustic tokenizers,” *ICML*, 2022.

[2] Mikhael Gromov, “Hyperbolic groups,” in *Essays in group theory*. Springer, 1987.

[3] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky, “Hyperbolic image embeddings,” in *IEEE CVPR*, 2020.

[4] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger, “Umap: Uniform manifold approximation and projection,” *Journal of Open Source Software*, 2018.

[5] Aaron Asael Smith, Rui Li, and Zion Tsz Ho Tse, “Reshaping healthcare with wearable biosensors,” *Scientific Reports*, vol. 13, no. 1, pp. 4998, 2023.

[6] Dídac Surís, Ruoshi Liu, and Carl Vondrick, “Learning the predictability of the future,” in *CVPR*, 2021.

[7] Desai Karan et al., “Hyperbolic image-text representations,” in *ICML*, 2023.

[8] Pal Avik et al., “Compositional entailment learning for hyperbolic vision-language models,” *ICLR*, 2025.