ASK: Adaptive Self-improving Knowledge Framework for Audio Text Retrieval

Anonymous ACL submission

Abstract

The dominant paradigm for Audio-Text Retrieval (ATR) relies on mini-batch-based contrastive learning. This process is constrained by what we define as the Gradient Locality Bottleneck (GLB), where optimization is limited to in-batch contrasts. This restricts the model from leveraging out-of-batch knowledge and consequently hinders long-tail learning. While external knowledge-enhanced methods can alleviate the GLB, we identify a critical, unaddressed side effect: the Representation-Drift Mismatch (RDM), where a static knowledge base becomes progressively misaligned with the evolving model, turning guidance into noise. To address this dual challenge, we propose the **A**daptive **S**elf-improving **K**nowledge (ASK) framework, a model-agnostic, plug-and-play solution. ASK breaks the GLB via multi-grained knowledge injection, systematically mitigates RDM through dynamic knowledge refinement, and introduces a novel adaptive reliability weighting scheme to ensure consistent knowledge contributes to optimization. State-of-the-art performance on established benchmarks demonstrates the efficacy of our proposed ASK framework. Our code is available at https://anonymous.4open.science/r/Code-FGL10.

1 Introduction

Audio-Text Retrieval (ATR) learns a shared embedding space for audio and text (Mei et al., 2022; Yan et al., 2024). The dominant paradigm relies on dual-encoder architectures trained with contrastive objectives like the NT-Xent loss (Chen et al., 2020), which optimizes representations by exclusively contrasting samples within a mini-batch (Figure 1, left). However, the reliance on in-batch negatives is a well-recognized limitation, often failing to provide sufficiently hard negatives to effectively structure the embedding space (Robinson et al., 2021). Critically, this paradigm structurally prevents the model from leveraging any out-of-batch information, leaving the vast majority of the dataset's semantic knowledge untapped during each optimization step.

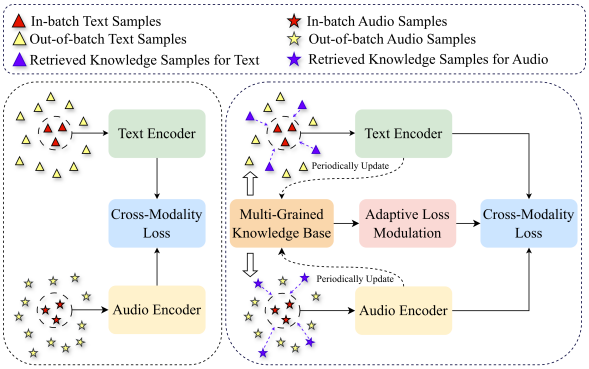

Figure 1: Comparison between the conventional batch-only paradigm (left) and our proposed **ASK** framework (right) with a periodically updated knowledge base and an adaptive loss modulation module.

In this work, we formalize this constraint as the Gradient Locality Bottleneck (GLB). We argue the GLB manifests in two critical failures: (1) it exacerbates semantic sparsity from under-specified text, as the model cannot access richer out-of-batch context to learn fine-grained acoustic details; and (2) it impairs long-tail generalization, a known challenge for contrastive methods (Kang et al., 2020), by preventing the model from forming robust decision boundaries for rare events.

A promising remedy is to augment training with an external knowledge base to access out-of-batch information (Khandelwal et al., 2019; Guu et al., 2020). However, this introduces a critical, unaddressed challenge: a Representation-Drift Mismatch (RDM) arises as the model's encoders evolve while the knowledge base remains static. The retrieved knowledge degrades from a source of semantic guidance to one of representational noise, destabilizing training and necessitating a co-

evolution of the model and its knowledge.

To systematically address this dual challenge, we propose the **A**daptive **S**elf-improving **K**nowledge (ASK) framework, a model-agnostic, plug-and-play solution (Figure 1, right). The ASK framework breaks the GLB by injecting information from a multi-grained knowledge base. To ensure the quality of this injection, a novel adaptive reliability weighting scheme modulates the final loss based on the cross-modal consistency of retrieved neighborhoods. Crucially, to prevent the knowledge from becoming stale, a dynamic refinement mechanism periodically updates the base, systemically mitigating RDM.

This synergistic design of reliability-governed injection and dynamic refinement proves highly effective. Extensive experiments show that ASK consistently and significantly outperforms strong baselines across multiple datasets, architectures, and interaction strategies, achieving new state-of-the-art performance.

Our main contributions are:

- We are the first to formally define the Gradient Locality Bottleneck (GLB) in contrastive learning and the consequent Representation-Drift Mismatch (RDM) in knowledge-enhanced methods, providing rigorous mathematical formalizations for both.
- We propose the **ASK** framework, a systematic solution to these challenges, featuring novel mechanisms for multi-grained knowledge injection, adaptive reliability weighting, and dynamic knowledge refinement.
- We demonstrate through extensive experiments that ASK achieves consistent state-of-the-art performance across diverse architectures and datasets, and validate the necessity of each component via comprehensive ablation studies.

2 Related Work

2.1 Feature Representations

Feature representation serves as the cornerstone of audio-text retrieval. Early Audio-Text Retrieval (ATR) systems relied on pairing handcrafted acoustic features like MFCCs (Huizen and Kurniati, 2021) with static word embeddings such as Word2Vec (Mikolov et al., 2013). The advent of deep learning has led to the adoption of powerful, pre-trained unimodal encoders. Text representations are now predominantly extracted from large language models like BERT (Devlin et al., 2019), while audio features are derived from deep models pre-trained on large-scale audio datasets, such as PANNs (Kong et al., 2020) and AST (Gong et al., 2021). More recently, the field has shifted towards large-scale cross-modal pre-training. Models like CLAP (Elizalde et al., 2022; Guzhov et al., 2021) leverage contrastive learning on vast audio-text datasets to directly learn a shared embedding space, significantly enhancing zero-shot capabilities. Our work builds upon these advanced encoders, proposing a novel mechanism to further enhance their representations during downstream fine-tuning.

2.2 Cross-Modal Interaction and Alignment

Cross-modal interaction is key to achieving semantic alignment in ATR. Early and prevalent approaches perform this at a global, sentence-level, using contrastive learning to align the final embeddings of entire audio clips and text descriptions (Radford et al., 2021; Wu et al., 2021; Mei et al., 2022). To capture more fine-grained relationships, recent works have focused on local, token-level interactions. These methods typically employ attention mechanisms or cross-modal Transformers to model correspondences between audio frames and text tokens (Lee et al., 2018; Lu et al., 2019; Xie et al., 2024; Yin et al., 2025). Our ASK framework is orthogonal to these design choices; it operates on the representations themselves and can be seamlessly integrated with both global and local interaction architectures.

3 Problem Formulation and Analysis

3.1 Preliminaries

In a standard Audio-Text Retrieval framework, a dual-encoder architecture, comprising an audio encoder $f_\theta(\cdot)$ and a text encoder $g_\phi(\cdot)$, maps an audio-text pair (a_i, t_i) to L2-normalized embeddings u_i and v_i. The encoders are optimized via a symmetric NT-Xent loss (Chen et al., 2020) over a mini-batch B. For a single view, the loss is:

$$\mathcal{L}_i = -\log \frac{\exp(u_i^\top v_i/\tau)}{\sum_{v_j \in B} \exp(u_i^\top v_j/\tau)} \qquad (1)$$

where τ is a temperature hyperparameter. Crucially, as shown in Eq. 1, the contrastive denominator is computed exclusively over samples within the mini-batch B. This inherent structural confinement is the direct cause of the bottleneck we analyze next.

3.2 The Gradient Locality Bottleneck

The batch-centric nature of Eq. 1 creates a fundamental limitation. To formalize this, we define the Out-of-Batch Influence (OBI) as the expected gradient norm of the loss $\mathcal{L}_B$ with respect to all out-of-batch embeddings:

$$\text{OBI}(\mathcal{L}_B) = \mathbb{E}_{k \in D \setminus B} \left[\left\| \frac{\partial \mathcal{L}_B}{\partial u_k} \right\|_2 + \left\| \frac{\partial \mathcal{L}_B}{\partial v_k} \right\|_2 \right] \tag{2}$$

A training paradigm suffers from a Gradient Locality Bottleneck (GLB) if its OBI is identically zero, indicating no gradient flow from out-of-batch data.

For the standard contrastive loss, $\mathcal{L}_B$ is exclusively a function of in-batch embeddings $\{u_j, v_j\}_{j \in B}$. Therefore, the partial derivatives with respect to any out-of-batch embedding u_k or v_k (where $k \notin B$) are necessarily zero. This directly results in $\text{OBI}(\mathcal{L}_B) = 0$, proving that standard ATR is strictly constrained by the GLB and cannot leverage the vast semantic information present in out-of-batch data.

3.3 The Representation Drift Mismatch

A direct approach to break the GLB is to perform knowledge injection, where out-of-batch knowledge is retrieved and fused with the current samples. This, however, introduces a critical challenge if the knowledge base remains static. A Representation Drift Mismatch (RDM) arises as the model's encoders at step t evolve away from the parameters used to build the knowledge base at step t_k.

To formalize this, we define RDM as the KL divergence (Kullback and Leibler, 1951) between the ideal neighborhood distribution P_{ideal} and the actual distribution P_{actual}. The ideal distribution is computed over a hypothetically up-to-date knowledge base, while the actual distribution uses the stale one:

$$\begin{aligned} P_{\text{ideal}}(j|i) &= \text{softmax}_j(\text{sim}(f_{\theta_t}(a_i), f_{\theta_t}(a_j))) \\ P_{\text{actual}}(j|i) &= \text{softmax}_j(\text{sim}(f_{\theta_t}(a_i), f_{\theta_{t_k}}(a_j))) \end{aligned} \tag{3}$$

The total RDM is then the expectation of this divergence over the dataset:

$$\text{RDM}(t, t_k) = \mathbb{E}_{a_i \in D} \left[D_{KL} \left(P_{\text{ideal}}(\cdot|i) \,||\, P_{\text{actual}}(\cdot|i) \right) \right] \tag{4}$$

As the time difference $\Delta t = t - t_k$ grows, RDM increases. This corrupts the training gradients by causing a deviation in the fused knowledge vectors, $\Delta\mathcal{K} = \mathcal{K}_{\text{actual}} - \mathcal{K}_{\text{ideal}}$, where each knowledge vector $\mathcal{K}$ is the average of the Top-K retrieved embeddings.

A larger deviation in the knowledge vector $\Delta\mathcal{K}$ directly translates to a greater potential deviation in the final parameter gradients. We provide a formal proof of this entire causal chain in Appendix A. The derivation first establishes the link between the knowledge deviation $\Delta\mathcal{K}$ and the gradient deviation, and then leverages Pinsker's inequality (Cover, 1999) to bound $\|\Delta\mathcal{K}\|_2$ with the RDM, establishing the key relationship:

$$\|\Delta\mathcal{K}\|_2 \leq C\sqrt{2 \cdot \text{RDM}(t, t_k)} \tag{5}$$

where C is a bounded constant. Eq. 5 proves that a higher RDM widens the potential error margin for the gradient, establishing a formal link to training instability and motivating our dynamic refinement mechanism.

4 The Adaptive Self-improving Knowledge Framework

In this section, we elaborate on each component of our proposed framework ASK, whose architecture is shown in Figure 2.

4.1 Formulation of Knowledge Bases

Our framework's first step is to construct multi-grained knowledge bases from a source dataset, $\mathcal{D}_k$. The choice of source is flexible; in our experiments, we explore three types to demonstrate versatility: 1) In-Domain $^+$: the training set itself, 2) Out-of-Domain †: WavCaps (Mei et al., 2024), and 3) Enriched In-Domain *: training set re-annotated by Gemini 2.5 (Comanici et al., 2025). From a chosen source, we construct two complementary bases.

Fine-Grained Knowledge Base. The fine-grained base, K_f, captures instance-level semantic details. It is formed by encoding all audio-text pairs in the source $\mathcal{D}_k = \{(a_j^k, t_j^k)\}_{j=1}^{N_k}$ using the current model encoders $f_\theta(\cdot)$ and $g_\phi(\cdot)$. The result is a collection of L2-normalized embedding pairs:

$$\begin{aligned} K_f &= \{(u_j^k, v_j^k)\}_{j=1}^{N_k}, \\ \text{where } u_j^k &= f_\theta(a_j^k), v_j^k = g_\phi(t_j^k) \end{aligned} \tag{6}$$

Coarse-Grained Knowledge Base. The coarse-grained base, K_c, provides a global semantic prior by storing a set of learned prototypes. These prototypes are generated by first partitioning the fine-grained embeddings via K-Means clustering into

N_c groups, and then distilling the salient features from each group. For the m-th audio cluster $\mathcal{C}_m^u$, which contains all member embeddings $\{u_j^k\}$, its prototype c_m^u is computed via max-pooling:

$$c_m^u = \text{MaxPool}(\{u_j^k \mid u_j^k \in \mathcal{C}_m^u\}) \tag{7}$$

An identical procedure is applied to the text embeddings to yield text prototypes $\{c_m^v\}_{m=1}^{N_c}$. The final coarse-grained base is the set of these prototype pairs, $K_c = \{(c_m^u, c_m^v)\}_{m=1}^{N_c}$.

4.2 Multi-Grained Knowledge Injection

With the knowledge bases established, we perform two parallel injection processes to create distinct fine-grained and coarse-grained enhanced embeddings for each training sample.

For the fine-grained injection, we first retrieve the Top-K nearest neighbors for a given embedding (e.g., audio u_i) from K_f, yielding the neighborhood set $\mathcal{N}_f(u_i)$. The retrieved embeddings are averaged to form a knowledge vector $\bar{u}_i^f$, which is then interpolated with the original embedding u_i:

$$u'_{i,f} = \rho u_i + (1-\rho)\bar{u}_i^f,$$
$$\text{where } \bar{u}_i^f = \frac{\sum_{(u_j^k, v_j^k)\in\mathcal{N}_f(u_i)} u_j^k}{K} \tag{8}$$

where ρ is an interpolation hyperparameter. An identical, parallel process is performed using the coarse-grained base K_c to produce the coarse-grained enhanced representation, $u'_{i,c}$. A symmetric procedure is applied to the text embedding v_i, ultimately yielding two distinct sets of enhanced embedding pairs for the final optimization: $(u'_{i,f}, v'_{i,f})$ and $(u'_{i,c}, v'_{i,c})$.

Breaking the Gradient Locality Bottleneck. This injection mechanism breaks the GLB (Sec. 3.2) by creating a gradient pathway to out-of-batch knowledge. For any out-of-batch knowledge item u_k^k retrieved by an in-batch sample u_i, its gradient is non-zero. Let $\mathcal{S}_k = \{i \in B \mid u_k^k \in \mathcal{N}_f(u_i)\}$ be the set of in-batch samples that retrieved u_k^k. The gradient of the loss $\mathcal{L}'_B$ w.r.t. u_k^k is:

$$\frac{\partial \mathcal{L}'_B}{\partial u_k^k} = \sum_{i\in\mathcal{S}_k} \frac{\partial \mathcal{L}'_B}{\partial u'_{i,f}} \frac{\partial u'_{i,f}}{\partial u_k^k} \tag{9}$$

From Eq. 8, the second partial derivative is a non-zero constant $\frac{1-\rho}{K}$. Since the first derivative is also non-zero, the total gradient is non-zero. Consequently, the OBI, defined in Eq. 2, becomes strictly positive: $\text{OBI}(\mathcal{L}'_B) > 0$. This quantitatively proves that our injection process breaks the GLB.

4.3 Adaptive Reliability Weighting

To mitigate the risk of injecting noisy knowledge from equally-weighted neighbors (Sec. 4.2), we introduce an adaptive weighting mechanism. This mechanism is based on the principle of cross-modal consistency: for a well-aligned audio-text pair (u_i, v_i), the neighborhoods retrieved by u_i and v_i should themselves be semantically consistent. We quantify this consistency to compute a reliability score for each neighbor, which in turn modulates its contribution to the final objective.

Fine-Grained Reliability Weighting. For each pair (u_i, v_i), we consider two fine-grained neighborhoods: the audio-retrieved audio set $\mathcal{U}_r = \{u_l^k\}_{l=1}^K$ and the text-retrieved audio–text set $\mathcal{N}_f(v_i) = \{(u_j^{k'}, v_j^k)\}_{j=1}^K$. We first assign each neighbor in $\mathcal{N}_f(v_i)$ a consistency score $\bar{s}_j$, defined as its average similarity to the audio-retrieved neighborhood:

$$\bar{s}_j = \frac{1}{K}\sum_{l=1}^{K} (u_j^{k'})^\top u_l^k. \tag{10}$$

These scores are subsequently normalized via a softmax function to yield the reliability weights $\mathbf{w}_f = \{w_j\}_{j=1}^K$:

$$w_j = \frac{\exp(\bar{s}_j)}{\sum_{m=1}^K \exp(\bar{s}_m)}. \tag{11}$$

The reliability-aware knowledge potential is then computed as the weighted similarity between u_i and the audio components of $\mathcal{N}_f(v_i)$:

$$\Psi_{i,f}^{T\to A} = \sum_{j=1}^{K} w_j \cdot \exp(u_i^\top u_j^{k'}). \tag{12}$$

A symmetric construction produces the text-side potential $\Psi_{i,f}^{A\to T}$, based on the audio-retrieved text neighborhood.

Coarse-Grained Reliability Weighting. An identical procedure is applied to the coarse-grained neighborhoods to produce the coarse-grained potentials, $\Psi_{i,c}^{T\to A}$ and $\Psi_{i,c}^{A\to T}$. These potentials represent the model's alignment with reliable, high-level semantic prototypes.

The resulting four reliability-aware potentials are core components that will be directly incorporated into our final optimization objective, as detailed in Section 4.5.

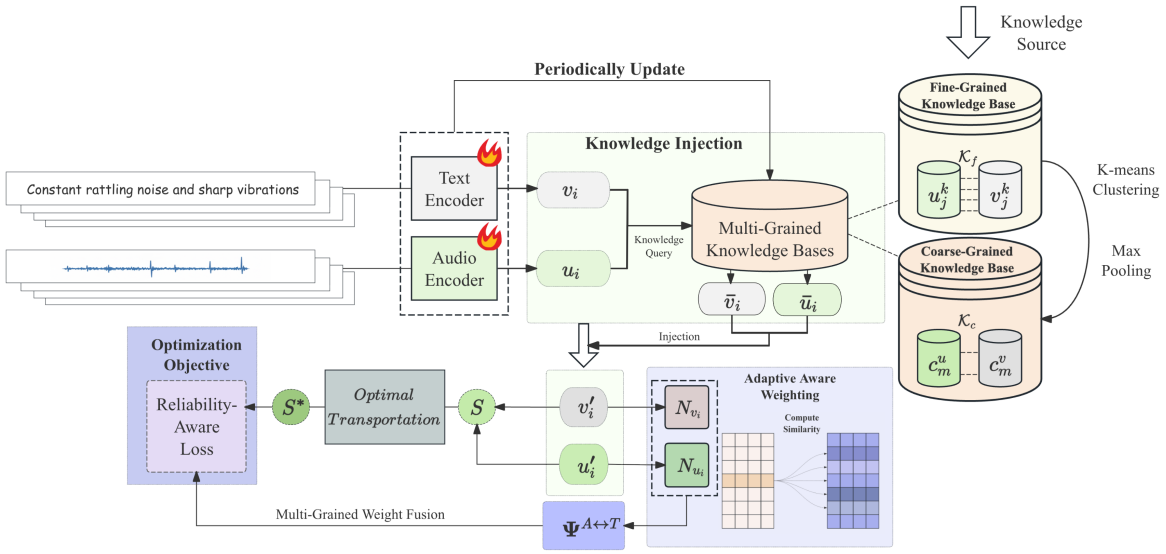

Figure 2: The proposed ASK framework. A multi-grained knowledge base (K_f, K_c) is periodically updated to mitigate RDM. During training, knowledge is injected into samples $(u_i \rightarrow u_i')$, and a cross-modal reliability weight (Ψ) is computed. A final loss is optimized using both an OT-realigned similarity matrix (S^*) and the reliability weight Ψ.

4.4 Dynamic Knowledge Refinement

As shown in Section 3.3, a static knowledge base leads to Representation Drift Mismatch (RDM), which induces increasing gradient misalignment during training. To mitigate this, we employ a dynamic refinement mechanism that periodically reconstructs the knowledge bases K_f and K_c using the current encoders. The update period $\mathcal{T}$ specifies the number of epochs between successive reconstructions.

This procedure directly controls the RDM. At each update step t, refinement sets the knowledge-base timestamp to $t_k = t$, making the ideal and actual neighborhood distributions identical, $P_{\text{ideal}} \equiv P_{\text{actual}}$. Thus, the RDM (Eq. 4) is reset to its minimum value:

$$\text{RDM}(t, t) = \mathbb{E}[D_{KL}(P_{\text{ideal}} \,\|\, P_{\text{ideal}})] = 0. \quad (13)$$

By periodically driving the RDM to zero, the mechanism also resets the upper bound on gradient deviation (Eq. 5), ensuring stable optimization and enabling the knowledge base to co-evolve with the model.

4.5 Unified Optimization Objective

The final optimization objective is constructed in two main stages. First, we compute NT-Xent losses on similarity matrices that have been realigned via Optimal Transport. Second, these losses are modulated by our reliability-aware knowledge potentials to form the final composite objective.

Loss on OT-Realigned Similarities. The process begins with the knowledge-enhanced embeddings from Section 4.2. For a mini-batch, we compute a fine-grained similarity matrix $\mathbf{S}_f$ and a coarse-grained one $\mathbf{S}_c$. Since the audio and text knowledge are retrieved independently, the distributions of their nearest neighbors within the batch may differ. To reconcile this potential discrepancy and find a globally optimal batch-level matching, we employ Optimal Transport (OT) (Cuturi, 2013) to learn an optimal transport plan $\mathbf{Q}^*$ (the full formulation is detailed in Appendix C). This plan is then used to produce the realigned similarity matrices $\mathbf{S}_f^*$ and $\mathbf{S}_c^*$:

$$\mathbf{S}_f^* = \left((1-\beta)\mathbf{I} + \beta\,\mathbf{Q}^*\right)\mathbf{S}_f \quad (14)$$

An identical process is applied to $\mathbf{S}_c$. Based on these realigned matrices, we define two NT-Xent loss components. The text-to-audio loss, $\mathcal{L}_{T\rightarrow A}$, is the sum of the fine- and coarse-grained objectives:

$$\mathcal{L}_{T\rightarrow A} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp((\mathbf{S}_f^*)_{ii}/\tau)}{\sum_{j=1}^{B}\exp((\mathbf{S}_f^*)_{ij}/\tau)} - \frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp((\mathbf{S}_c^*)_{ii}/\tau)}{\sum_{j=1}^{B}\exp((\mathbf{S}_c^*)_{ij}/\tau)}. \quad (15)$$

The audio-to-text loss, $\mathcal{L}_{A\to T}$, is formulated symmetrically.

Reliability-Aware Objective. The OT-realigned losses above do not yet account for the cross-modal consistency of the retrieved knowledge. To incorporate this, we use the knowledge potentials computed in Section 4.3 as reliability modulators. We first define the reliability-aware terms, e.g., for the text-to-audio direction:

$$\mathcal{F}_f^{T\to A} = \frac{1}{|B|}\sum_{i=1}^{|B|} -\log \Psi_{i,f}^{T\to A}$$
$$\mathcal{F}_c^{T\to A} = \frac{1}{|B|}\sum_{i=1}^{|B|} -\log \Psi_{i,c}^{T\to A} \quad (16)$$

The final text-to-audio loss, $\mathcal{L}^*_{T\to A}$, is then the base OT-realigned loss, modulated by a weighted sum of these reliability terms:

$$\mathcal{L}^*_{T\to A} = (1+\lambda_f \mathcal{F}_f^{T\to A}+\lambda_c \mathcal{F}_c^{T\to A})\cdot\mathcal{L}_{T\to A} \quad (17)$$

where λ_f and λ_c are hyperparameters. The final audio-to-text loss, $\mathcal{L}^*_{A\to T}$, is computed symmetrically. The overall loss for the ASK framework is the average of these two modulated objectives:

$$\mathcal{L}_{\text{ASK}} = \frac{1}{2}(\mathcal{L}^*_{T\to A} + \mathcal{L}^*_{A\to T}) \quad (18)$$

This composite objective ensures the model learns from multi-grained knowledge that is both globally aligned at the batch level and weighted by its cross-modal reliability. Furthermore, we provide a theoretical proof in Appendix B that demonstrates the convergence properties of our ASK framework.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We evaluate our method on two standard benchmarks: AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020). Following prior work (Mei et al., 2022; Xie et al., 2024; Yan et al., 2024), we report audio-to-text (A2T) and text-to-audio (T2A) retrieval performance using Recall at K (R@K, for K=1, 5, 10).

Baselines. To validate the model-agnostic nature of ASK, we integrate it into two types of baselines. **1) Global Interaction:** We use a PANNs-based ResNet-38 (Kong et al., 2020) + BERT (Devlin et al., 2019) pair (Mei et al., 2022), and a ViT-based CED-Base (Dinkel et al., 2024) + SONAR-TE (Duquenne et al., 2023) pair, following ML-CLAP (Yan et al., 2024) but trained only on English. **2) Local Interaction:** We adapt the GPA (Xie et al., 2024) setup, using its ResNet-38 + BERT architecture but removing the Sinkhorn inference module to form the baseline, and set the same maximum number of tokens for the entire dataset.

Implementation Details. All models are trained with the Adam optimizer (Adam et al., 2014). The ResNet-BERT architecture is trained for 50 epochs on AudioCaps (batch size 32) and Clotho (batch size 24), with an initial learning rate of 5×10^{-5}, which is decayed by a factor of 10 every 20 epochs. The CED-SONAR models are trained for 10 epochs with a decay step applied every 4 epochs. We use the Faiss library (Douze et al., 2025) for efficient neighbor search. Unless specified otherwise, the hyperparameters for our ASK framework are set as follows: we retrieve $K = 10$ neighbors, with a coarse-grained prototype set of size $N_c = 512$. The knowledge injection ratio is $\rho = 0.2$, and the OT-realignment factor is $\beta = 0.2$. The reliability modulation weights are $\lambda_f = 0.2$ and $\lambda_c = 0.3$. The knowledge base is dynamically refined every $\mathcal{T} = 15$ epochs. All experiments were conducted on 2 NVIDIA A100 and 8 RTX 4090 GPUs.

5.2 Main Results

We evaluate the effectiveness of our proposed ASK framework by integrating it into various baseline models and comparing their performance on the AudioCaps and Clotho datasets. The results are organized by the cross-modal interaction strategy.

Global Interaction Strategy. Table 1 presents the results for models employing a global, sentence-level interaction strategy. Our ASK framework demonstrates substantial and consistent improvements across both datasets and architectures. When applied to the ResNet-BERT baseline on AudioCaps, ASK improves the A2T R@1 score by a remarkable 6.0% absolute and the T2A R@1 score by 3.2% absolute . This strong performance gain validates the effectiveness of our core mechanisms in breaking the GLB and mitigating RDM. Furthermore, ASK proves to be model-agnostic, delivering significant gains on the more powerful transformer-based CED-SONAR architecture as well. For instance, on the challenging Clotho dataset, it boosts the A2T R@1 by up to 1.7% absolute and the T2A R@1 by 1.4% absolute. The results also highlight the flexibility of ASK in leveraging diverse

Table 1: Results for Audio-Text-Retrieval on AudioCaps and Clotho under the global interaction strategy. The symbols $^{+}$, †, and * denote the use of knowledge from the original training set, WavCaps, and the Gemini-annotated training set, respectively.

Method	AudioCaps						Clotho					
	A2T			T2A			A2T			T2A		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Architecture: ResNet-38 + BERT												
Mei et al., 2022	36.3	68.6	81.5	32.2	68.2	81.2	16.3	39.1	51.5	14.2	37.3	49.9
ASK†	**42.3**	73.3	84.2	34.6	69.6	82.9	17.3	40.2	**54.1**	14.8	38.1	50.7
ASK*	39.5	73.2	85.3	34.2	69.1	81.9	**18.5**	40.1	53.6	14.7	38.3	50.1
ASK^{+}	42.0	**74.2**	**85.4**	**35.4**	**70.2**	**83.1**	17.5	**40.3**	**54.1**	**15.2**	**38.5**	**51.1**
Architecture: CED-Base + SONAR-TE												
Yan et al., 2024	39.6	69.8	81.7	31.9	69.2	82.8	18.0	39.5	53.0	14.9	39.9	53.1
ASK†	**43.3**	73.7	84.4	34.8	70.6	84.0	19.0	41.5	56.5	**16.3**	40.3	**55.4**
ASK*	41.9	**74.1**	**85.6**	**34.9**	**70.9**	**84.1**	18.5	41.6	56.9	16.0	40.6	55.1
ASK^{+}	40.9	71.6	84.3	33.7	70.3	83.5	**19.7**	**43.3**	**57.3**	16.0	**41.5**	**55.2**

knowledge sources, with different sources showing strengths on different dataset-architecture combinations.

Local Interaction Strategy. We also validate ASK on a strong baseline with a local, token-level interaction strategy (Xie et al., 2024). The Audio-to-Textretrieval results are presented in Table 2. The full results, including the Text-to-Audio retrieval scores, are detailed in Appendix D.

The results demonstrate that ASK delivers consistent and significant gains even on this fine-grained architecture. On AudioCaps, our best variant, ASK*, improves the R@1 score by a substantial margin of 2.6% absolute. On the more challenging Clotho dataset, ASK^{+} achieves the top R@1 performance, boosting the baseline by 1.4% absolute. These improvements underscore the universal benefit of our framework; breaking the GLB and mitigating RDM are crucial enhancements regardless of whether the model's interaction mechanism is global or local.

Zero-Shot Generalization. In addition to the in-domain evaluations, we conduct a challenging zero-shot cross-dataset experiment to further assess the generalization capabilities of ASK. The results, detailed in Appendix E, demonstrate that ASK significantly improves the model's performance when transferring from AudioCaps to Clotho, confirming its strong generalization benefits.

Table 2: Results for Audio-to-Text Retrieval under the local interaction strategy. The symbols $^{+}$, †, and * denote different knowledge sources in Section 4.1.

Method	AudioCaps			Clotho		
	R@1	R@5	R@10	R@1	R@5	R@10
Xie et al., 2024	41.1	73.8	85.2	18.1	40.2	53.4
ASK†	42.9	75.1	86.4	19.1	**41.9**	53.9
ASK*	**43.7**	**75.8**	86.2	19.2	41.6	**54.5**
ASK^{+}	43.1	74.0	**86.9**	**19.5**	41.4	**54.5**

5.3 Ablation Study and Analysis

To validate the contribution of each component within our ASK framework, we conduct a series of ablation studies on the AudioCaps dataset using the ResNet-BERT architecture and an in-domain knowledge source. The results are presented in Table 3.

Impact of Multi-Grained Knowledge Bases. We first analyze the necessity of our multi-grained design. Removing the fine-grained knowledge base results in a substantial performance drop of 4.3% absolute in A2T R@1, confirming the critical role of instance-level details for precise retrieval. Similarly, removing the coarse-grained base leads to a 4.6% drop in A2T R@1, which underscores the importance of the global semantic prior provided by the prototypes. The model, which leverages both, significantly outperforms either single-granularity variant, demonstrating that the fine- and coarse-grained knowledge sources are complementary.

Table 3: Ablation experiments on AudioCaps dataset using the ResNet-38 + BERT architecture. $^+$ denotes the utilization of knowledge derived from AudioCaps training set.

G.	Method	A2T			T2A		
		R@1	R@5	R@10	R@1	R@5	R@10
	w/o ASK (baseline)	36.3	68.6	81.5	32.2	68.2	81.2
1	w/o Fine-grained Knowledge Base	37.7	70.4	81.8	31.9	67.3	81.0
	w/o Coarse-grained Knowledge Base	37.4	67.6	81.3	31.2	66.6	81.0
2	w/o the Knowledge Injection Step	39.1	72.7	84.1	34.5	69.1	82.6
	w/o OT Alignment Correction	41.1	73.4	85.2	34.2	69.4	82.8
3	w/o Adaptive Reliability Weighting	39.3	72.2	83.6	33.9	68.9	81.6
4	w/o the Dynamic Knowledge Refinement	39.2	71.0	83.8	34.1	68.7	81.5
	Our Full **ASK**$^+$	**42.0**	**74.2**	**85.4**	**35.4**	**70.2**	**83.1**

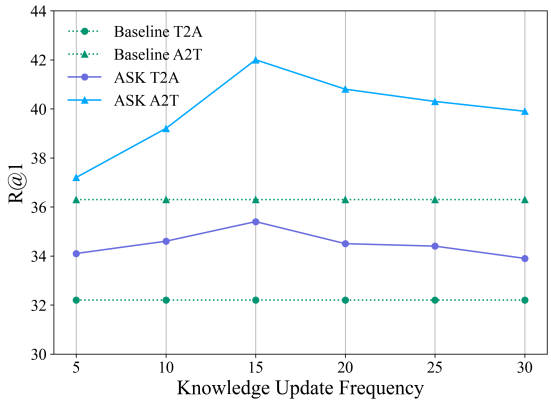

Figure 3: Ablation experiment on **ASK**$^+$. Effect of the number $\mathcal{T}$ of Knowledge Update.

Impact of Core ASK Mechanisms. We then ablate the core mechanisms of ASK. **1) Knowledge Injection:** Disabling the knowledge injection step causes a notable drop of 2.9% in A2T R@1. This empirically validates that creating gradient pathways to out-of-batch data is the primary driver for breaking the GLB and enhancing representations. **2) Reliability Weighting:** Ablating our adaptive reliability weighting mechanism results in a significant 2.7% drop in A2T R@1 and a 1.5% drop in T2A R@1. This provides strong evidence that not all retrieved knowledge is equally beneficial, and that modulating the loss based on cross-modal consistency is crucial for mitigating the impact of noises and achieving robust performance.

Impact of Dynamic Knowledge Refinement. We evaluate the effect of the knowledge-base update period $\mathcal{T}$ on mitigating RDM. As shown in Table 3, disabling dynamic refinement leads to a 2.8% drop in A2T R1, empirically validating our theoretical claim in Section 3.3 that unchecked RDM introduces stale and misaligned knowledge.

Figure 3 shows that performance improves as the update frequency increases, reaching an optimum at $T = 15$ epochs, which surpasses both the static knowledge base and the baseline. However, overly frequent updates degrade performance, indicating a trade-off: while frequent updates curtail RDM, they can also destabilize the knowledge representation before the model fully adapts. These findings highlight the necessity of a co-evolving knowledge base and careful tuning of the update frequency.

6 Conclusion

In this paper, we identified and formalized two fundamental challenges in knowledge-enhanced Audio-Text Retrieval: the Gradient Locality Bottleneck, which confines standard contrastive learning to mini-batches, and the consequent Representation-Drift Mismatch, which arises from using static knowledge bases with evolving models. To address this dual challenge, we proposed the Adaptive Self-improving Knowledge framework. ASK is a model-agnostic, plug-and-play solution that breaks the GLB via multi-grained knowledge injection, mitigates RDM through dynamic knowledge refinement, and ensures reliability with a novel adaptive weighting scheme. Extensive experiments demonstrate that ASK consistently and significantly improves performance across diverse architectures and datasets, achieving new state-of-the-art results.

7 Limitations

While our experiments demonstrate the broad effectiveness of ASK, we observed that using a significantly larger and more diverse out-of-domain knowledge source (WavCaps) did not always yield proportionally larger gains compared to the in-domain source on our test sets. We hypothesize that this may be due to the relatively close data distribution between the training and test sets of benchmarks like AudioCaps and Clotho. In such scenarios, injecting knowledge from a vastly different distribution might introduce a slight domain shift that tempers the benefits of increased diversity.

References

Kingma DP Ba J Adam and 1 others. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6).

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. 2024. Ced: Consistent ensemble distillation for audio tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 291–295. IEEE.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.

Bernardo Elizalde and 1 others. 2022. Clap: Contrastive language–audio pretraining. *arXiv preprint arXiv:2206.04769*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Andrey Guzhov, Felix Raue, Joern Hees, and Andreas Dengel. 2021. Audioclip: Extending clip to image, text and audio. *arXiv preprint arXiv:2106.13043*.

Roy Rudolf Huizen and Florentina Tatrin Kurniati. 2021. Feature extraction with mel scale separation method on noise audio recordings. *arXiv preprint arXiv:2112.14930*.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. 2020. Exploring balanced feature spaces for representation learning. In *International conference on learning representations*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. 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)*, pages 119–132.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image–text matching. *arXiv preprint arXiv:1803.08024*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations. *arXiv preprint arXiv:1908.02265*.

Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang. 2022. On metric learning for audio-text cross-modal retrieval. *arXiv preprint arXiv:2203.15537*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Alec Radford and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. 2021. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986.

Bing Su and Gang Hua. 2017. Order-preserving wasserstein distance for sequence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1057.

Ho-Hsiang Wu, Bernardo Elizalde, and Zeyu Wang. 2021. Wav2clip: Learning robust audio representations from clip. *arXiv preprint arXiv:2110.11499*.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. 2024. Gpa: Global and prototype alignment for audio-text retrieval. In *Proc. Interspeech*, volume 2024, pages 5078–5082.

Zhiyong Yan, Heinrich Dinkel, Yongqing Wang, Jizhong Liu, Junbo Zhang, Yujun Wang, and Bin Wang. 2024. Bridging language gaps in audio-text retrieval. *arXiv preprint arXiv:2406.07012*.

Yuguo Yin, Yuxin Xie, Wenyuan Yang, Dongchao Yang, Jinghan Ru, Xianwei Zhuang, Liming Liang, and Yuexian Zou. 2025. Atri: Mitigating multilingual audio text retrieval inconsistencies by reducing data distribution errors. *arXiv preprint arXiv:2502.14627*.

A Derivation and Visualization of RDM's Impact

This appendix provides a detailed derivation of the relationship between the Representation Drift Mismatch (RDM) and training stability. The core premise of RDM is that a model's representation space is non-stationary during training. We first provide a visualization in Figure 4 that empirically demonstrates this phenomenon. It shows how the embeddings of the same audio clips, encoded by a model without dynamic updates, drift significantly as training progresses. Our goal in the following sections is to formally prove that this observed drift leads to a greater potential for gradient misalignment.

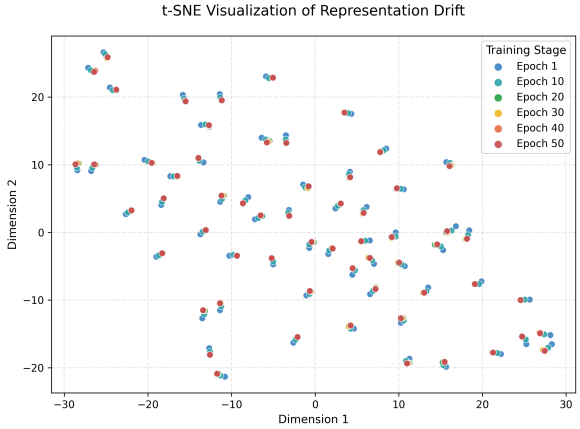

Figure 4: t-SNE visualization of Representation Drift. Embeddings of a fixed set of audio samples, encoded by the same model at different training epochs, are plotted. The progressive shift in embedding positions (from Epoch 1 [blue] to Epoch 50 [red]) empirically validates the core premise of RDM: a static knowledge base becomes misaligned with the non-stationary representation space over time.

Gradient Formulation. We consider a simplified loss function $\mathcal{L} = \mathcal{L}_{\text{main}}(u_i, u_i')$ that incorporates a knowledge-enhanced representation $u_i' = (1-\rho)u_i + \rho\mathcal{K}$, where $u_i = f_{\theta_t}(a_i)$ and $\mathcal{K}$ is the expected representation of retrieved knowledge. The gradient of the loss with respect to the model parameters θ_t is:

$$\nabla_{\theta_t}\mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial u_i} + (1-\rho)\frac{\partial \mathcal{L}}{\partial u_i'}\right)\frac{\partial u_i}{\partial \theta_t} \tag{19}$$

Linking Gradient Deviation to Knowledge Deviation. The difference between the ideal gradient ($\nabla_{\theta_t}\mathcal{L}_{\text{ideal}}$) and the actual gradient ($\nabla_{\theta_t}\mathcal{L}_{\text{actual}}$)

arises from the difference in their respective knowledge vectors, $\mathcal{K}_{\text{ideal}}$ and $\mathcal{K}_{\text{actual}}$. Let the gradient difference vector be $\Delta\nabla = \nabla_{\theta_t}\mathcal{L}_{\text{actual}} - \nabla_{\theta_t}\mathcal{L}_{\text{ideal}}$. This difference is primarily driven by the change in the loss derivative term $\frac{\partial\mathcal{L}}{\partial u'_i}$.

To analyze this relationship, we use a first-order Taylor expansion of the loss gradient term around the ideal representation u'_{ideal}. The difference can be approximated as:

$$\frac{\partial\mathcal{L}_{\text{actual}}}{\partial u'_i} - \frac{\partial\mathcal{L}_{\text{ideal}}}{\partial u'_i} \approx H_{\mathcal{L}}(u'_{\text{ideal}}) \cdot (u'_{\text{actual}} - u'_{\text{ideal}}) \tag{20}$$

where $H_{\mathcal{L}}$ is the Hessian matrix of the loss function with respect to its input. Since $u'_{\text{actual}} - u'_{\text{ideal}} = \rho(\mathcal{K}_{\text{actual}} - \mathcal{K}_{\text{ideal}}) = \rho\Delta\mathcal{K}$, we can see that the deviation in the loss gradient is approximately proportional to the deviation in the knowledge vector:

$$\Delta\nabla \propto H_{\mathcal{L}} \cdot \Delta\mathcal{K} \tag{21}$$

This establishes a direct relationship: a larger deviation in the fused knowledge vector $\Delta\mathcal{K}$ leads to a larger deviation in the final parameter gradient $\Delta\nabla$. The next step is therefore to bound the magnitude of $\Delta\mathcal{K}$ using the RDM.

Bounding the Knowledge Deviation via RDM. We now bound the norm of the deviation $\|\Delta\mathcal{K}\|_2$ using the RDM. We leverage Pinsker's inequality (Cover, 1999), which relates the KL divergence to the Total Variation Distance (D_{TV}):

$$\begin{aligned} D_{TV}(P_1, P_2) &= \frac{1}{2}\sum_j |P_1(j) - P_2(j)| \\ &\leq \sqrt{\frac{1}{2}D_{KL}(P_1 \,||\, P_2)} \end{aligned} \tag{22}$$

Applying this to our distributions gives $D_{TV}(P_{\text{ideal}}, P_{\text{actual}}) \leq \sqrt{\frac{1}{2}\text{RDM}(t, t_k)}$. We can then bound $\|\Delta\mathcal{K}\|_2$:

$$\begin{aligned} \|\Delta\mathcal{K}\|_2 &= \|\sum_j (P_{\text{actual}}(j) - P_{\text{ideal}}(j))z_j\|_2 \\ &\leq \sum_j |P_{\text{actual}}(j) - P_{\text{ideal}}(j)|\|z_j\|_2 \\ &\leq \left(\max_j \|z_j\|_2\right) \cdot 2 \cdot D_{TV}(P_{\text{ideal}}, P_{\text{actual}}) \\ &\leq C\sqrt{2 \cdot \text{RDM}(t, t_k)} \end{aligned} \tag{23}$$

where $C = \max_j \|z_j\|_2$ is a bounded constant.

Conclusion. Combining these steps, we have established a formal link: an increase in RDM (Eq. 4) widens the upper bound on the knowledge vector deviation $\|\Delta\mathcal{K}\|_2$ (Eq. 23), which in turn increases the potential magnitude of the gradient deviation $\Delta\nabla$ (Eq. 20). This increases the risk of gradient misalignment, which can lead to training instability. Our dynamic knowledge refinement mechanism is designed to mitigate this risk by periodically resetting the RDM to zero.

B Theoretical Justification and Convergence of the ASK Objective

In this section, we provide a theoretical justification for the ASK framework. We demonstrate that our training procedure can be viewed as a principled alternating optimization algorithm designed to maximize the log-likelihood of the observed data, which in turn guarantees the monotonic non-increase of our final loss function and thus ensures convergence.

Probabilistic Formulation with Latent Knowledge. The primary goal of Audio-Text Retrieval is to find model parameters θ^* that maximize the log-likelihood of observing matched audio-text pairs $x_i = (a_i, t_i)$:

$$\theta^* = \max_\theta \mathcal{L}(\theta) = \max_\theta \sum_i \log p(x_i; \theta) \tag{24}$$

We conceptualize our approach by introducing latent variables, $z_i = (z_{i,f}, z_{i,c})$, representing the unobserved "optimal" knowledge for each sample x_i. The observed data likelihood is the marginal likelihood over these latent variables:

$$p(x_i; \theta) = \sum_{z_i} p(x_i, z_i; \theta) \tag{25}$$

Thus, the optimization objective becomes:

$$\theta^* = \max_\theta \sum_i \log \sum_{z_i} p(x_i, z_i; \theta) \tag{26}$$

The summation inside the logarithm makes direct optimization intractable.

Deriving the Evidence Lower Bound. To create a tractable objective, we introduce an arbitrary distribution $Q(z_i)$ and apply Jensen's Inequality to derive a lower bound on the log-likelihood, known as the Evidence Lower Bound (ELBO), denoted as

$\mathcal{F}(Q, \theta)$:

$$\log p(x_i;\theta) = \log \sum_{z_i} Q(z_i) \frac{p(x_i, z_i;\theta)}{Q(z_i)}$$
$$\geq \sum_{z_i} Q(z_i) \log \frac{p(x_i, z_i;\theta)}{Q(z_i)} \quad (27)$$
$$\mathcal{F}(Q, \theta_i) = \mathbb{E}_{Q(z_i)}[\log p(x_i, z_i;\theta)] \quad (28)$$
$$- \mathbb{E}_{Q(z_i)}[\log Q(z_i)]$$

Maximizing $\log p(x_i;\theta)$ is achieved by iteratively maximizing this lower bound $\mathcal{F}$ with respect to Q and θ.

The ASK Framework as an Alternating Optimization Algorithm. Let θ_t be the parameters at iteration t. The ASK training process alternates between two stages.

Stage 1: Auxiliary Distribution Update. In this stage, we fix θ_t and approximate the optimal auxiliary distribution $Q_t(z_i)$ which should be the true posterior $p(z_i|x_i;\theta_t)$. We assume independence between fine- and coarse-grained knowledge: $Q_t(z_i) = Q_{t,f}(z_{i,f})Q_{t,c}(z_{i,c})$.

- The retrieval of Top-K neighbors defines the support of $Q_{t,f}$ and $Q_{t,c}$.
- We define the probability mass of these distributions over a specific neighbor z_j using our reliability weights (Eq. 11):

$$Q_{t,f}(z_{i,f} = z_j) := w_{j,f}(\theta_t),$$
$$Q_{t,c}(z_{i,c} = z_j) := w_{j,c}(\theta_t) \quad (29)$$

Stage 2: Model Parameter Update. In this stage, we fix Q_t and maximize the ELBO with respect to θ, which is equivalent to maximizing $\mathbb{E}_{Q_t}[\log p(x_i, z_i;\theta)]$. We model the joint log-probability as a sum of independent fine- and coarse-grained components, e.g., for the text-to-audio direction:

$$\log p(x_i, z_i;\theta) \approx \left(-\mathcal{L}_{OT,f}(\theta) - \log \Psi_{i,f}^{T\leftarrow A}(\theta)\right)$$
$$+ \left(-\mathcal{L}_{OT,c}(\theta) - \log \Psi_{i,c}^{T\leftarrow A}(\theta)\right)$$
$$+ \left(-\log Z(\theta)\right) \quad (30)$$

where $Z(\theta)$ is a normalization constant. The maximization objective is to minimize the negative expectation of this log-probability under Q_t. Substituting Eq. 29 and Eq. 30, this objective becomes:

$$\mathcal{L}_m = -\sum_i \mathbb{E}_{Q_t(z_i)}[\log p(x_i, z_i;\theta)]$$
$$\approx \sum_i (\mathbb{E}_{Q_{t,f}}[\mathcal{L}_{OT,f} + \log \Psi_{i,f}] \quad (31)$$
$$+ \mathbb{E}_{Q_{t,c}}[\mathcal{L}_{OT,c} + \log \Psi_{i,c}])$$

Our final modulated loss from Eq. 17,

$$\mathcal{L}^*_{T\to A} = (1+\lambda_f \mathcal{F}_f^{T\to A} + \lambda_c \mathcal{F}_c^{T\to A}) \cdot \mathcal{L}_{T\to A} \quad (32)$$

where $\mathcal{F} = -\log \Psi$, is a principled and sophisticated implementation of this maximization objective. Minimizing $\mathcal{L}_{\text{ASK}}$ effectively performs this parameter update.

Proof of Convergence. This two-stage alternating optimization guarantees that the total objective is non-decreasing at each full iteration, $\mathcal{L}(\theta_{t+1}) \geq \mathcal{L}(\theta_t)$. Consequently, minimizing the negative log-likelihood guarantees that the loss is monotonically non-increasing. Given that $\mathcal{L}_{\text{ASK}}$ is bounded below by zero, the Monotone Convergence Theorem ensures that the sequence of loss values converges to a limit, and the parameters $\{\theta_t\}$ converge to a stationary point

C Optimal Transport for Batch-level Alignment

This section details the entropy-regularized Optimal Transport (OT) formulation used to refine the batch-wise similarity matrices. Given a batch of knowledge-enhanced pairs, we compute a similarity matrix, e.g., the fine-grained matrix $\mathbf{S}_f \in \mathbb{R}^{B\times B}$. We then seek an optimal transport plan $\mathbf{Q} \in \mathbb{R}^{B\times B}$, where $\mathbf{Q}_{ij}$ represents the soft-alignment probability between the i-th text and the j-th audio. The optimal plan $\mathbf{Q}^*$ is found by solving the following regularized optimization problem:

$$\mathbf{Q}^* = \max_{\mathbf{Q}\in\mathcal{C}} \langle \mathbf{Q}, \mathbf{S}_f \rangle + \varepsilon H(\mathbf{Q})$$
$$\text{s.t. } \mathcal{C} = \{\mathbf{Q} \in \mathbb{R}^{B\times B} \mid \mathbf{Q}\mathbf{1}_B = \boldsymbol{\mu},\ \mathbf{Q}^\top \mathbf{1}_B = \boldsymbol{\nu}\}, \quad (33)$$

where $\langle \mathbf{Q}, \mathbf{S}_f \rangle = \text{tr}(\mathbf{Q}^\top \mathbf{S}_f)$ is the total similarity score. $H(\mathbf{Q}) = -\sum_{i,j} \mathbf{Q}_{ij} \log \mathbf{Q}_{ij}$ is the entropy regularizer, controlled by $\varepsilon > 0$. The constraints enforce that the marginals of $\mathbf{Q}$ must sum to predefined distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, which represent the importance of each instance. Following prior work (Su and Hua, 2017), we set both $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ to a uniform distribution over the batch, i.e., $\frac{1}{|B|}\mathbf{1}_{|B|}$. This

problem is efficiently solved for the optimal plan $\mathbf{Q}^*$ using the Sinkhorn-Knopp algorithm (Cuturi, 2013).

D Full Results for Local Interaction Strategy

This section provides the complete retrieval results for our experiments on the local, token-level interaction baseline, including both Audio-to-Text and Text-to-Audio directions. Table 4 presents the full comparison.

Table 4: Full results for Audio-Text Retrieval on AudioCaps and Clotho under the local interaction strategy. The symbols $^+$, †, and * denote different knowledge sources in Section 4.1.

Audio-to-Text						
	AudioCaps			Clotho		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Xie et al., 2024	41.1	73.8	85.2	18.1	40.2	53.4
ASK†	42.9	75.1	86.4	19.1	**41.9**	53.9
ASK*	**43.7**	**75.8**	86.2	19.2	41.6	**54.5**
ASK$^+$	43.1	74.0	**86.9**	**19.5**	41.4	**54.5**
Text-to-Audio						
	AudioCaps			Clotho		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Xie et al., 2024	34.1	70.0	82.2	15.1	37.9	50.2
ASK†	34.5	**71.1**	**83.1**	16.2	38.5	51.3
ASK*	34.6	70.5	82.7	**16.3**	38.4	51.5
ASK$^+$	**35.1**	70.8	**83.1**	16.0	**38.8**	**52.1**

As demonstrated in Table 4, ASK consistently improves upon the baseline in the Text-to-Audio retrieval direction as well. On AudioCaps, ASK$^+$ achieves the highest R@1 score, improving the baseline by 1.0% absolute. On Clotho, the ASK* variant delivers the strongest R@1 performance with a significant gain of 1.2% absolute. These results confirm that the benefits of our proposed mechanisms are symmetric, enhancing both retrieval directions and validating the overall effectiveness of the ASK framework on fine-grained architectures.

E Zero-Shot Generalization

To further assess the generalization capabilities of our ASK framework, we conduct a zero-shot cross-dataset evaluation. In this setup, models are trained exclusively on the AudioCaps training set and then directly evaluated on the Clotho test set, without any fine-tuning. This challenging setting tests the model's ability to generalize to a different data distribution. The results for the global ResNet-BERT architecture are presented in Table 5.

Table 5: Zero-shot generalization performance on the Clotho test set. All models were trained only on AudioCaps. The symbols $^+$, †, and * denote different knowledge sources in Section 4.1.

	A2T			T2A		
Method	R@1	R@5	R@10	R@1	R@5	R@10
Mei et al., 2022	12.8	29.0	39.7	10.1	27.6	38.3
ASK†	**14.1**	30.3	**43.7**	**11.9**	**31.2**	**42.8**
ASK*	13.6	30.2	40.5	**11.9**	30.5	40.3
ASK$^+$	13.6	**31.6**	43.3	11.5	30.8	42.6

The results demonstrate that ASK significantly enhances zero-shot generalization. Notably, ASK†, leveraging the large-scale out-of-domain WavCaps dataset, achieves the best performance with absolute gains of 1.3% in A2T R@1 and 1.8% in T2A R@1. This indicates that exposing the model to diverse external knowledge fosters robust semantic representations, enabling effective transfer to the unseen Clotho dataset. These findings validate ASK as a framework for genuine generalization rather than simple in-domain memorization.

F Stability Analysis

Table 6: Detailed retrieval performance on AudioCaps with standard deviations. The symbols †, *, and + denote different knowledge sources in Section 4.1.

	A2T			T2A		
Method	R@1	R@5	R@10	R@1	R@5	R@10
ASK†	**42.3**$_{\pm 0.3}$	73.3$_{\pm 0.8}$	84.2$_{\pm 0.6}$	34.6$_{\pm 0.5}$	69.6$_{\pm 0.4}$	82.9$_{\pm 0.9}$
ASK*	39.5$_{\pm 0.3}$	73.2$_{\pm 0.4}$	85.3$_{\pm 0.6}$	34.2$_{\pm 0.6}$	69.1$_{\pm 0.7}$	81.9$_{\pm 0.3}$
ASK$^+$	42.0$_{\pm 0.2}$	**74.2**$_{\pm 0.5}$	**85.4**$_{\pm 0.6}$	**35.4**$_{\pm 0.3}$	**70.2**$_{\pm 0.3}$	**83.1**$_{\pm 0.7}$

In our experiments, we evaluated various architectures and interaction mechanisms. To ensure reliability, all results reported in the main text are averages derived from multiple independent runs. In this section, we provide a detailed stability analysis focusing on the ResNet-38 + BERT architecture under the global interaction strategy. Table 6 presents the specific statistics for this setting. The results demonstrate minimal fluctuation, confirming that our method maintains high stability and statistical significance.