SARA: STABLE ANCHORS AND RIGOROUS ALIGNMENT FOR FINE-GRAINED AUDIO-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained Audio-Text Retrieval (ATR) requires precise token-level semantic alignment, yet existing prototype-based methods face two coupled limitations. First, without geometric constraints, learned prototypes collapse into redundant global representations, reducing fine-grained matching to global alignment. Second, mini-batch contrastive training limits the supply of hard negatives for distinguishing subtle semantic differences. We propose Stable Anchors and Rigorous Alignment (SARA) to address both jointly. To suppress prototype collapse and preserve inter-prototype diversity, SARA replaces dynamic pooling with learnable queries, applies orthogonal regularization to the resulting prototypes, and uses optimal transport to balance cross-modal prototype assignment so that every prototype contributes to matching. To further train these stabilized prototypes to distinguish fine-grained semantic differences, SARA mines hard negatives from a momentum-updated dynamic queue, focusing contrastive training on high-similarity negative audio-text pairs. Experiments on AudioCaps and Clotho under the unified encoder configuration show that SARA achieves the highest mean R@1 in both retrieval directions. On AudioCaps, SARA improves mean R@1 over the baseline (GPA) by 6.31 and 3.05 percentage points for audio-to-text and text-to-audio retrieval, respectively. Diagnostic analysis and case studies further show that SARA effectively mitigates prototype collapse and supports fine-grained semantic alignment.

1 INTRODUCTION

Audio-Text Retrieval (ATR) (Elizalde et al., 2023; Guzhov et al., 2022) bridges acoustic signals and natural language, enabling users to retrieve relevant sounds from audio collections through free-form text descriptions (Koepke et al., 2022). Beyond early coarse-grained alignment between global embeddings, recent work shifts toward fine-grained alignment (Wu et al., 2025; Fang et al., 2025), which decomposes auditory scenes into discrete events and maps them to specific linguistic tokens for better interpretability and precision. Yet, despite progress made by prototype-based methods such as GPA (Xie et al., 2024), two limitations remain at the representation and optimization levels.

Specifically, at the representation level, prevailing methods generate prototypes via dynamic pooling (Xie et al., 2024; Fu et al., 2025) without explicitly constraining directional correlations among prototypes. Consequently, different prototypes converge toward the same direction, resulting in a phenomenon we call *Prototype Collapse* (Jing et al., 2021). Under complete collapse, fine-grained matching degenerates into global alignment and loses the ability to distinguish local correspondences. Even when prototype collapse is avoided, a challenge remains at the optimization level. Insufficient fine-grained discriminative signals during training limit the model's ability to distinguish semantically similar audio-text candidates. However, standard mini-batch contrastive training restricts negatives to the current batch, reducing opportunities to encounter high-similarity candidates and limiting the training signals available for fine-grained discrimination (Kalantidis et al., 2020; Fu et al., 2025).

To address both challenges, we propose Stable Anchors and Rigorous Alignment (SARA). For prototype collapse at the representation level, we introduce *Stable Anchors* to preserve prototype diversity through geometric constraints and support balanced cross-modal matching. Specifically, we use a set of learnable queries shared across samples to replace dynamic pooling and construct semantic

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.

prototypes, while applying orthogonal regularization to the resulting prototypes to suppress directional redundancy and preserve inter-prototype diversity. Building on this, we introduce optimal transport (Cuturi, 2013) to constrain the total assignment mass on both the audio and text sides so that every prototype participates in cross-modal matching. At the optimization level, we address insufficient fine-grained discriminative signals through *Rigorous Alignment*, training the stabilized prototypes to distinguish semantically similar samples. To this end, we maintain a momentum-updated dynamic queue that expands the negative candidate pool beyond the current mini-batch, mine high-similarity negatives from it, and construct additional negatives through interpolation. These hard negatives focus contrastive training on subtle semantic differences, providing geometrically diverse prototypes with the training signals needed for fine-grained discrimination. Our main contributions are listed as follows:

- We propose *Stable Anchors*, combining learnable-query prototype construction, orthogonal constraints on the resulting prototypes, and optimal transport matching to suppress prototype collapse, preserve inter-prototype diversity, and achieve balanced cross-modal prototype assignment.
- We incorporate momentum-queue hard negative mining and negative interpolation into *Rigorous Alignment* to provide training signals for fine-grained discrimination. Through this training process, we train the prototypes to distinguish semantically similar audio-text candidates.
- Under the unified encoder configuration, SARA achieves the highest mean R@1 in both retrieval directions on AudioCaps and Clotho. On AudioCaps, SARA improves mean R@1 over GPA by 6.31 and 3.05 percentage points for audio-to-text and text-to-audio retrieval, respectively. Through diagnostic analysis and case studies, we further show that SARA mitigates prototype collapse and supports fine-grained semantic alignment.

2 Preliminaries and Problem Analysis

We first formalize the prototype generation paradigm in Audio-Text Retrieval (ATR), then analyze two problems faced by existing methods: the geometric phenomenon of prototype collapse and the optimization bottleneck caused by the scarcity of informative negatives in mini-batch training.

2.1 Standard Prototype Generation

Given an audio-text pair (A, T), we extract feature sequences via their respective encoders. Let the audio features be $X^a \in \mathbb{R}^{L_a \times d}$ and the text features be $X^t \in \mathbb{R}^{L_t \times d}$, where the superscripts a, t denote the audio and text modalities, respectively. Here, L_a, L_t are the sequence lengths, d is the feature dimension, and x_i^a denotes the i-th audio token feature. Prototype learning methods pool variable-length features into K compact semantic prototypes $P^a = \{p_1^a, \ldots, p_K^a\} \in \mathbb{R}^{K \times d}$, typically via dynamic pooling based on linear projection (Ilse et al., 2018; Xie et al., 2024; Xin et al., 2024):

$$p_k^a = \sum_{i=1}^{L_a} \alpha_{k,i} x_i^a, \quad \text{where } \boldsymbol{\alpha} = \text{Softmax}(f_{proj}(X^a)). \tag{1}$$

Here, $\alpha_{k,i}$ is the pooling weight assigned to token i when aggregating prototype k, and f_{proj} is a linear layer that computes pooling scores. Text prototypes are generated symmetrically. Each prototype p_k is intended to capture specific local semantics, thereby enabling fine-grained cross-modal alignment.

2.2 The Phenomenon of Prototype Collapse

Fine-grained matching requires different prototypes to capture distinct local semantics. However, we observe directional redundancy among dynamically pooled prototypes within each modality. We term the convergence of different prototypes toward the same direction *Prototype Collapse*, in which pairwise cosine similarities approach unity such that $\cos(p_i, p_j) \to 1$ for $i \neq j$.

Under complete collapse, all nonzero prototypes within a modality share the same direction and can be written as $p_k = c_k u$, where u is their common unit direction and $c_k > 0$ is the magnitude of each

prototype. The rows of the prototype matrix are then positive scalar multiples of one another, giving $\mathrm{rank}(P) = 1$.

This directional redundancy directly causes multi-prototype matching to degenerate. Consider complete collapse in both modalities,

$$p_k^a = c_k^a u^a, \qquad p_k^t = c_k^t u^t, \tag{2}$$

where u^a and u^t are the common unit directions of the audio and text prototypes, respectively, and $c_k^a, c_k^t > 0$. All cross-modal prototype pairs have the same cosine similarity. For matching weights satisfying $w_k \geq 0$ and $\sum_{k=1}^{K} w_k = 1$, the prototype matching score becomes

$$S_{fine} = \sum_{k=1}^{K} w_k \cos(p_k^a, p_k^t) = \sum_{k=1}^{K} w_k \langle u^a, u^t \rangle = \langle u^a, u^t \rangle. \tag{3}$$

Although the model retains K prototypes, matching depends on only one direction from each modality. Distinctions among local correspondences disappear, and the multi-prototype branch degenerates into parameter-redundant global matching. We further present prototype attention distributions through GPA case studies in Appendix G, providing visual evidence to complement the analysis of prototype collapse.

2.3 Scarcity of Informative Negatives

Preserving prototype diversity addresses directional redundancy, while fine-grained retrieval further requires distinguishing audio-text candidates with similar content. Hard negatives are negative samples with high matching scores against the anchor.

For an anchor with one positive sample and a set of negatives, let S_{pos} and S_{neg} denote the positive and negative similarities, respectively, and let $\tau > 0$ denote the temperature. Under InfoNCE (Oord et al., 2018), the loss derivative with respect to a negative similarity is

$$\frac{\partial \mathcal{L}}{\partial S_{neg}} = \frac{1}{\tau} \frac{\exp(S_{neg}/\tau)}{\exp(S_{pos}/\tau) + \sum_{neg'} \exp(S_{neg'}/\tau)}. \tag{4}$$

Within the same candidate set, higher-similarity negatives have larger score derivatives, giving them greater weight in contrastive optimization. Standard mini-batch training restricts the available negatives to the current batch. When highly similar negatives are scarce, training provides few comparisons between audio-text candidates with similar content. Expanding the negative pool and selecting high-similarity candidates increases opportunities for these comparisons, motivating a scalable supply of informative hard negatives.

2.4 Summary of Analysis

The analysis establishes two design goals for fine-grained audio-text alignment.

- **Representation.** Suppress Prototype Collapse and preserve distinct prototype directions so that multi-prototype matching can distinguish different local correspondences.
- **Optimization.** Supply high-similarity negatives that concentrate contrastive optimization on difficult matching decisions and support fine-grained discrimination.

These goals correspond to Stable Anchors and Rigorous Alignment in SARA. Stable Anchors construct diverse prototype representations, while Rigorous Alignment trains discriminative cross-modal matching through informative negatives.

3 Methodology

To address the representation and optimization problems analyzed in Sec. 2, we propose SARA. Through *Stable Anchors*, we preserve prototype diversity and achieve balanced cross-modal matching, as detailed in Secs. 3.1 and 3.2. Through *Rigorous Alignment*, we focus contrastive training on hard negatives to learn fine-grained semantic differences, as detailed in Sec. 3.3.

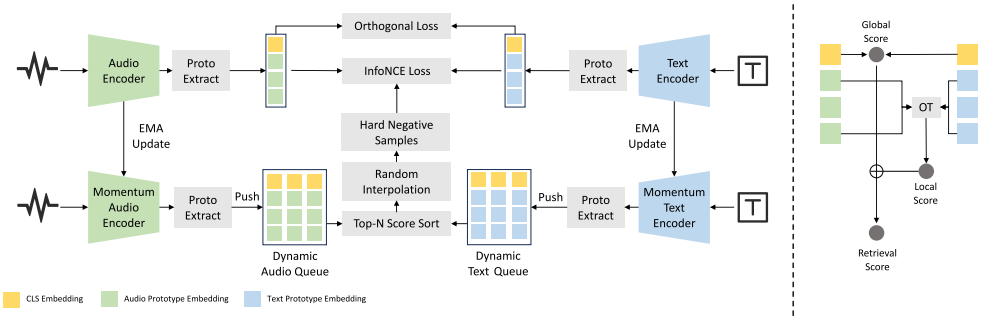


Figure 1: **The architecture of the proposed SARA framework. Left:** The training process integrates queue-based hard negative mining for contrastive optimization while employing an orthogonal loss to regularize prototypes against collapse. **Right:** Computation of global and fine-grained similarities and their combination into the retrieval score.

3.1 Stable Anchors via Orthogonal Prototype

To suppress prototype collapse, we construct prototypes through learnable queries shared across samples and apply orthogonal regularization to the resulting prototypes. The former provides consistent prototype construction across samples, while the latter constrains directional correlations among prototypes to preserve their diversity.

From Dynamic Pooling to Stable Anchors. We learn separate sets of K queries for the audio and text modalities. Within each modality, the query matrix $Q \in \mathbb{R}^{K\times d}$ is shared across all samples and updated through retrieval training. Given an input feature sequence $X \in \mathbb{R}^{L\times d}$, we aggregate the input through cross-attention to form the prototype matrix $P \in \mathbb{R}^{K\times d}$:

$$P = \text{Attention}(Q, X, X). \tag{5}$$

Here, Q provides the queries, while X provides the keys and values. Shared query parameters provide a consistent prototype construction across samples, while the attention weights select content from the current input.

Orthogonal Regularization. To suppress Prototype Collapse, we directly regularize directional correlations among the generated prototypes. Let $\hat{P}$ denote the row-normalized prototype matrix with $\hat{p}_k = p_k/\|p_k\|_2$. The orthogonal loss is defined as

$$\mathcal{L}_{ortho} = \|\hat{P}\hat{P}^T - I\|_F^2 = \sum_{i\neq j}\langle \hat{p}_i, \hat{p}_j\rangle^2. \tag{6}$$

This loss penalizes correlations between distinct normalized prototypes, driving their pairwise inner products toward zero and suppressing the directional redundancy of Prototype Collapse.

3.2 Balanced Assignment with Optimal Transport

Building on the prototype diversity preserved within each modality in Sec. 3.1, we use optimal transport to balance cross-modal prototype assignment. Specifically, we constrain the total assignment mass on both the audio and text sides and use Sinkhorn-Knopp iterations to compute the matching plan.

Balanced Prototype Matching. Through orthogonal regularization in Eq. 6, we constrain prototype directions within each modality, but leave the allocation of cross-modal matching weights unconstrained. Normalizing these weights in only one direction still allows multiple prototypes to allocate most of their weight to the same counterpart, producing many-to-one matching. We therefore impose equal marginal constraints on both sides so that every audio and text prototype participates in matching with the same total assignment mass.

Sinkhorn Alignment. For normalized audio and text prototypes, we use cosine distance as the matching cost, defined by $C_{ij} = 1 - \langle \hat{p}_i^a, \hat{p}_j^t \rangle$. We obtain a nonnegative transport plan $\Gamma \in \mathbb{R}^{K \times K}$ through the entropy-regularized objective

$$\Gamma^* = \arg\min_{\Gamma \geq 0} \sum_{i,j} \Gamma_{ij} C_{ij} - \epsilon H(\Gamma), \tag{7}$$

where $H(\Gamma) = -\sum_{i,j} \Gamma_{ij} \log \Gamma_{ij}$ is the entropy and $\epsilon > 0$ controls its regularization strength. We impose the constraints $\Gamma \mathbf{1} = \mathbf{1}$ and $\Gamma^T \mathbf{1} = \mathbf{1}$. Every audio and text prototype thus receives equal total assignment mass, with cosine costs determining how that mass is distributed across prototype pairs. Entropy regularization supports differentiable matching through continuous assignment weights.

We use Sinkhorn-Knopp iterations to alternate row and column normalization and approximately solve this objective. Denoting the computed transport plan by Γ, the fine-grained matching score is

$$S_{fine} = \sum_{i,j} \Gamma_{ij} \cos(p_i^a, p_j^t). \tag{8}$$

3.3 Rigorous Alignment via Hard Negative Mining

Stable Anchors preserve geometric diversity among prototypes, while *Rigorous Alignment* further trains the model to distinguish audio-text samples with similar content. To this end, we mine hard negatives from a momentum-updated dynamic queue and densify negative sampling through interpolation, concentrating contrastive training on difficult matching decisions.

Momentum Queue and Hard Negative Mining. We maintain a FIFO momentum queue for each modality, storing M historical representations with $M \gg B$, where B is the batch size. This expands the negative pool beyond the current mini-batch. For each anchor q, we rank queue candidates from the opposite modality by their matching scores and select the Top-N pool $\{z_n\}_{n=1}^N$. Selecting multiple high-similarity candidates concentrates training on difficult matches and provides pairs for subsequent interpolation.

Interpolation for Boundary Densification. The Top-N pool provides discrete hard negatives. Following hard negative mixing (Kalantidis et al., 2020), we generate S additional negatives by linearly interpolating between randomly sampled pairs z_i, z_j from this pool.

$$z'_{neg} = \lambda z_i + (1 - \lambda) z_j, \quad \lambda \sim U(0, 1). \tag{9}$$

These interpolated representations densify negative sampling between mined candidates, providing additional comparisons for fine-grained discrimination.

3.4 Optimization Objective

Similarity Scoring. The final cross-modal similarity combines global and fine-grained scores.

$$S(A, T) = S_{global}(A, T) + S_{fine}(A, T). \tag{10}$$

Here, S_{global} is the cosine similarity between [CLS] tokens and S_{fine} is the OT-based prototype alignment score. This combines global context matching with fine-grained event alignment.

InfoNCE with Densified Negatives. We define the candidate set for the i-th anchor as $\mathcal{U}_i = \mathcal{B} \cup \mathcal{H}_i$, where $\mathcal{B}$ contains all text samples in the current mini-batch and $\mathcal{H}_i$ contains the N mined negatives and the S interpolated negatives generated from them. Taking the audio-to-text direction as an example, the loss is

$$\mathcal{L}_i^{a2t} = -\log \frac{\exp(S(A_i, T_i)/\tau)}{\sum_{T_j \in \mathcal{U}_i} \exp(S(A_i, T_j)/\tau)}. \tag{11}$$

Here, τ is the temperature. The text-to-audio loss $\mathcal{L}_i^{t2a}$ is computed symmetrically.

Total Loss. The final objective jointly minimizes the bi-directional InfoNCE loss and the orthogonal loss.

$$\mathcal{L}_{total} = \frac{1}{B} \sum_{i=1}^{B} \left(\mathcal{L}_i^{a2t} + \mathcal{L}_i^{t2a} \right) + \gamma \mathcal{L}_{ortho}. \tag{12}$$

Here, γ controls the strength of orthogonal regularization, suppressing repeated prototype directions while the model learns cross-modal matching.

4 Experimental Results

We evaluate SARA on AudioCaps and Clotho in terms of retrieval performance and prototype geometry. We first introduce the experimental setup and compare retrieval results under a unified encoder configuration. We then examine how prototype diversity and discriminative training affect retrieval through geometric diagnosis and component ablations. Finally, we analyze the choice of prototype number in relation to the distribution of events and relations in AudioCaps and illustrate prototype specialization through a case study.

4.1 Setup

Datasets. We evaluate the model on two benchmarks: AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020). AudioCaps contains approximately 50k audio clips with concise descriptions. In contrast, Clotho features longer, more complex descriptions and higher label noise, serving as a stress test for model robustness.

Evaluation Metrics. We evaluate retrieval performance using R@1 and R@5 in both directions and analyze audio and text prototype geometry using APCS and normalized effective rank ρ.

To measure directional similarity among prototypes, we compute average pairwise cosine similarity APCS by averaging pairwise cosine similarities within each prototype set and then across N sets, obtaining

$$\text{APCS} = \frac{1}{N}\sum_{n=1}^{N}\frac{2}{K(K-1)}\sum_{i<j}\cos(p_{n,i}, p_{n,j}), \tag{13}$$

where $p_{n,i}$ is the i-th prototype in the n-th set. APCS closer to 1 indicates more aligned prototype directions.

To further measure the concentration of singular values in the prototype matrix, we compute normalized effective rank ρ from the entropy of normalized singular values (Roy & Vetterli, 2007; Jing et al., 2021) as

$$\rho = \frac{\exp(H(\sigma)) - 1}{K - 1}, \qquad H(\sigma) = -\sum_{i}\bar{\sigma}_i \log \bar{\sigma}_i, \tag{14}$$

where σ_i are singular values, $\bar{\sigma}_i = \sigma_i / \sum_j \sigma_j$, and zero-valued terms contribute zero to the entropy. For $K > 1$, $\rho = 0$ when only one singular value is nonzero, whereas $\rho = 1$ when all K singular values are equal and nonzero.

Implementation Details. To ensure fair comparison with the baseline, we uniformly use CED-base (Dinkel et al., 2024) as the audio encoder and BERT (Devlin et al., 2019) as the text encoder. Key hyperparameters are: number of prototypes $K = 3$; momentum queue size $M = 7200$. For Hard Negative Mining, we select Top-N ($N = 10\% \cdot M$) candidates and interpolate $S = 50\% \cdot N$ samples. The orthogonal weight γ is set to 1.0 for AudioCaps and 1.5 for Clotho. In the Sinkhorn-Knopp algorithm, we set the number of iterations to 3. We use the Adam (Kingma, 2014) optimizer with an initial learning rate of 5×10^{-5} and multiply the learning rate by 0.1 every 5 epochs. All models are trained for 20 epochs on a single NVIDIA RTX 4090 GPU with a batch size of 24. All experiments are run three times with seeds $\in \{1997, 42, 2026\}$, and we report mean$\pm$std.

4.2 Main Results

As shown in Tab. 1, we compare SARA with Global methods including CLAP-LAION (Wu et al., 2023), Cacophony (Zhu et al., 2024), CLAP (Elizalde et al., 2023), and SupCLAP (Luo et al., 2025), and with Fine-grained methods including ASK (Fu et al., 2025), FLAM (Wu et al., 2025), THA-DCR (Xin et al., 2024), and the prototype baseline GPA (Xie et al., 2024). To ensure a strictly fair comparison, we rerun all external baselines on the same CED-base audio encoder and BERT text encoder as SARA, based on the official implementations where available and reproducing from the papers otherwise.

Table 1: Performance comparison with global and fine-grained methods on AudioCaps and Clotho datasets.

Method	AudioCaps				Clotho			
	A2T		T2A		A2T		T2A	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLAP-LAION	47.06±0.24	80.36±0.75	38.92±0.16	75.77±0.56	20.35±0.89	44.66±1.32	16.70±0.51	42.39±0.36
Cacophony	48.66±0.69	79.38±1.10	39.70±0.23	75.76±0.27	19.68±0.24	43.44±1.31	16.10±0.18	41.69±0.22
CLAP	49.77±0.69	80.50±0.52	39.54±0.35	76.21±0.27	20.64±0.80	44.37±1.33	16.25±0.33	42.12±0.25
SupCLAP	48.80±0.36	80.32±0.70	39.56±0.19	76.12±0.55	20.29±1.26	44.37±0.59	16.83±0.19	42.33±0.43
ASK	42.81±0.84	76.14±0.33	35.57±0.71	71.79±0.14	19.33±0.17	41.79±1.00	15.66±0.14	39.81±0.46
FLAM	49.74±0.58	80.70±0.67	38.03±0.26	75.34±0.52	20.10±0.44	45.22±1.02	15.90±0.31	**42.41±0.18**
THA-DCR	50.53±2.06	83.03±1.16	41.17±0.76	76.63±0.72	19.20±0.63	43.75±0.58	15.96±0.66	41.26±0.08
GPA (Baseline)	49.53±0.42	81.19±2.28	40.04±0.38	76.00±0.46	19.87±0.96	44.76±0.98	16.26±0.47	42.28±0.21
SARA (Ours)	**55.84±1.15**	**84.36±0.31**	**43.09±0.56**	**76.82±0.46**	**21.02±0.22**	**45.84±0.75**	**16.89±0.49**	**42.41±0.80**

Table 2: Diagnostic analysis of prototype geometry on AudioCaps. We compare retrieval performance (R@1) with geometric metrics (APCS and ρ) to diagnose collapse.

Model Setting	Retrieval R@1		Audio Geometry		Text Geometry	
	A2T	T2A	APCS ↓	ρ ↑	APCS ↓	ρ ↑
GPA (Baseline)	49.53±0.42	40.04±0.38	0.61±0.33	0.43±0.23	1.00±0.00	0.07±0.02
GPA w/ HN	52.14±1.29	**43.09±0.31**	1.00±0.00	0.12±0.01	1.00±0.00	0.07±0.02
SARA w/o HN queue	49.60±0.94	39.66±1.04	0.01±0.00	1.00±0.00	0.00±0.01	1.00±0.00
SARA (Full)	**55.84±1.15**	**43.09±0.56**	**0.21±0.01**	**0.96±0.01**	**0.17±0.02**	**0.98±0.01**

On AudioCaps, SARA achieves the highest mean R@1 and R@5 in both retrieval directions. A2T R@1 reaches 55.84%, exceeding GPA by 6.31 percentage points and THA-DCR by 5.31 percentage points. T2A R@1 reaches 43.09%, improving over GPA by 3.05 percentage points. SARA also achieves A2T and T2A R@5 of 84.36% and 76.82%, respectively, showing that its retrieval gains extend across both R@1 and R@5.

On Clotho, SARA again achieves the highest mean R@1 in both retrieval directions, reaching 21.02% for A2T and 16.89% for T2A. It also achieves the highest A2T R@5 of 45.84% and matches FLAM at the highest T2A R@5 of 42.41%. These results establish improvements over GPA in both R@1 and R@5 across the two datasets and retrieval directions under the same encoder configuration.

4.3 Diagnostic Analysis

We examine R@1 together with APCS and ρ in Tab. 2 to investigate the relationship between prototype geometry and retrieval performance.

Structural Failure of Dynamic Prototypes. GPA attains Text APCS of 1.00 and ρ of 0.07, indicating that its text prototypes collapse to nearly the same direction. Audio APCS and ρ are 0.61 and 0.43, respectively. In both audio case studies in Appendix G, we further observe Dead Prototypes (Van Den Oord et al., 2017) whose attention weights remain near zero across the temporal axis, providing sample-level evidence of prototype inactivity beyond the geometric metrics.

Exposure of Collapse under Discriminative Pressure. Introducing Hard Negative Mining raises GPA's A2T R@1 from 49.53% to 52.14% and T2A R@1 from 40.04% to 43.09%. Meanwhile, APCS reaches 1.00 in both modalities, Audio ρ drops to 0.12, and Text ρ remains at 0.07. The attention maps in Appendix G further show highly overlapping prototype patterns. These results demonstrate that hard negative mining improves retrieval without resolving Prototype Collapse.

Geometric Diversity and Retrieval Performance. SARA w/o HN queue, corresponding to $M = 0$ in Tab. 5, attains APCS near zero and ρ of 1.00 in both modalities. However, A2T and T2A R@1 remain close to GPA at 49.60% and 39.66%. Full SARA raises these results to 55.84% and 43.09%, respectively, while maintaining low APCS and ρ close to 1. This comparison shows that restoring

geometric diversity alone is insufficient to improve retrieval. With discriminative training, SARA improves retrieval while preserving prototype diversity.

4.4 Ablation Study

We evaluate the contributions of orthogonal regularization, OT, and the hard-negative pipeline through three component ablations in Tab. 3.

Table 3: Component-wise ablation on AudioCaps. Each row removes one component from full SARA; ΔR@1 reports the drop against the full model.

Variant	Audio-to-Text		Text-to-Audio	
	R@1	ΔR@1	R@1	ΔR@1
Full SARA (Ours)	**55.84±1.15**	—	**43.09±0.56**	—
w/o HN ($N = 0, S = 0$)	54.30±1.73	−1.54	42.86±0.79	−0.23
w/o OT (Softmax)	53.19±1.27	−2.65	42.43±0.39	−0.66
w/o Ortho ($\gamma = 0$)	52.94±0.75	−2.90	41.47±0.23	−1.62

Component Configurations. The w/o Ortho variant sets $\gamma = 0$ while retaining the learnable queries, OT, and the hard-negative pipeline. The w/o OT variant replaces Sinkhorn assignment with Softmax. The w/o HN variant retains the momentum queue but replaces hard negative mining with random sampling from the queue and disables interpolation. This differs from SARA w/o HN queue in Tab. 2, which removes the queue entirely.

Orthogonal Regularization and OT. Removing orthogonal regularization reduces A2T and T2A R@1 by 2.90 and 1.62 percentage points, respectively, the largest drops among these three ablations. The $\gamma = 0$ experiment in Appendix D also yields APCS near 1 in both modalities, showing that the retrieval decline accompanies Prototype Collapse. Replacing OT with Softmax reduces R@1 by 2.65 and 0.66 percentage points in the two directions. These results support the role of orthogonal regularization in preserving prototype diversity and the contribution of balanced OT assignment to retrieval. The orthogonal-weight sweep in Appendix D further shows that reducing APCS beyond the default setting does not improve R@1.

Hard-Negative Pipeline. With the momentum queue retained, removing mining and interpolation reduces mean A2T and T2A R@1 by 1.54 and 0.23 percentage points, respectively, indicating a larger benefit in A2T. Tab. 5 in the appendix further compares retrieval performance across queue sizes, mining budgets, and interpolation budgets.

4.5 Prototype Number and Event Density

We examine the choice of $K = 3$ through the event and relation structure of AudioCaps captions, with retrieval results across prototype counts reported in Tab. 5 in the appendix.

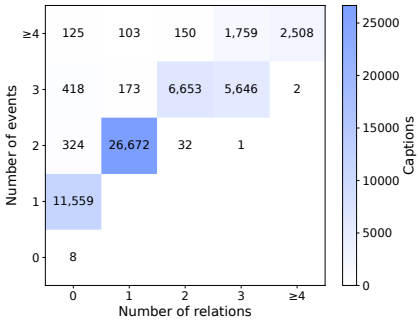


Figure 2: **Event and Relation Counts on AudioCaps.** Joint distribution of event count and relation count across the train, val, and test captions. The most frequent combination is two events and one relation, occurring in 26,672 captions.

Event Density and Prototype Configuration. The event and relation structure of AudioCaps captions provides a semantic motivation for $K = 3$. Fig. 2 shows the joint distribution of event and relation counts, where **2 Events + 1 Relation** is the most frequent combination, occurring in 26,672 captions. Three prototypes provide a compact configuration for separately representing the two events and their relation in this common compound-event structure. The annotation procedure and distributions by split are provided in Appendix C.

This configuration is also reflected in the learned prototype specialization. In the analyzed compound-event samples at $K = 3$, two text prototypes aggregate the content of the two events and the third attends to their relation, forming an **Entity A + Relation + Entity B** decomposition, while audio prototypes selectively attend to different temporal locations. Section 4.6 illustrates this specialization in a concrete example, with additional attention visualizations provided in Appendix I.

4.6 Case Study

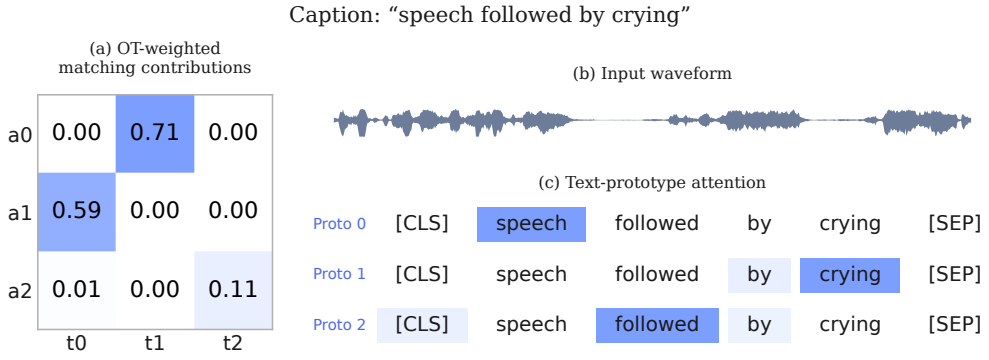


Figure 3: **Prototype specialization in SARA.** An AudioCaps example with the caption "speech followed by crying." **(a)** OT-weighted matching contributions $\Gamma_{ij} \cos(p_i^a, p_j^t)$. Rows index audio prototypes, and columns t0, t1, and t2 correspond to Proto 0, Proto 1, and Proto 2 in (c). **(b)** Input audio waveform. **(c)** Token attention for the three text prototypes. Shading indicates relative attention strength within each row.

Following the event-density analysis, we visualize an AudioCaps example in Fig. 3 to examine how the learned prototypes represent a compound-event caption. For "speech followed by crying," the displayed token attention shows that Proto 0 and Proto 1 attend to *speech* and *crying*, respectively, while Proto 2 emphasizes the connective *followed*. This case illustrates prototype specialization across events and their relation. Corresponding to this content selection, the two event-focused text prototypes receive their largest OT-weighted matching contributions from different audio prototypes. Appendix I provides additional prototype visualizations, while Appendix J further examines this example under hard-negative perturbations.

5 Conclusion

In this paper, we present SARA for fine-grained audio-text retrieval through Stable Anchors and Rigorous Alignment. Stable Anchors combine learnable-query prototype construction with orthogonal regularization and balanced cross-modal assignment through Optimal Transport. Rigorous Alignment strengthens discriminative training through hard negative mining and interpolation from a momentum queue. Under the unified encoder configuration, SARA achieves the highest mean R@1 in both retrieval directions on AudioCaps and Clotho, improving AudioCaps A2T R@1 over GPA by 6.31 percentage points. Our diagnosis further shows that retrieval gains can coexist with Prototype Collapse, while restoring geometric diversity alone is insufficient to improve retrieval. By jointly addressing the representation and optimization problems, SARA improves retrieval while maintaining geometrically diverse prototypes. Event-density analysis and qualitative cases further connect the compact prototype configuration to event and relation specialization, providing complementary evidence for understanding fine-grained alignment.

AI USE STATEMENT

In this work, we used generative AI tools for supporting qualitative data analysis of AudioCaps captions: the DeepSeek V4 API annotated acoustic events and the relations that connect them, which grounds the event density analysis. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, or interpreting experimental results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing the research paper to improve readability and for organizing the LaTeX source. We have reviewed all AI-assisted work. Event and relation labels produced by DeepSeek V4 were checked against the captions, and all proposed wording and LaTeX edits were checked by the authors against the experimental records before inclusion. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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

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.

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

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

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Minghui Fang, Shengpeng Ji, Jialong Zuo, Xize Cheng, Wenrui Liu, Xiaoda Yang, Ruofan Hu, Jieming Zhu, and Zhou Zhao. Gta: Towards generative text-to-audio retrieval via multi-scale tokenizer. In *Proc. Interspeech*, pp. 2650–2654, 2025.

Siyuan Fu, Xuchen Guo, Mingjun Liu, Hongxiang Li, Boyin Tan, Gongxi Zhu, Xianwei Zhuang, Jinghan Ru, Yuxin Xie, and Yuguo Yin. Ask: Adaptive self-improving knowledge framework for audio text retrieval. *arXiv preprint arXiv:2512.19703*, 2025.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. Compa: Addressing the gap in compositional reasoning in audio-language models. *arXiv preprint arXiv:2310.08753*, 2023.

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

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33: 21798–21809, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 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)*, pp. 119–132, 2019.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685, 2022.

Jiehui Luo, Yuguo Yin, Yuxin Xie, Jinghan Ru, Xianwei Zhuang, Minghua He, Aofan Liu, Zihan Xiong, and Dongchao Yang. Supclap: Controlling optimization trajectory drift in audio-text contrastive learning with support vector regularization. *arXiv preprint arXiv:2509.21033*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. Flam: Frame-wise language-audio modeling. *arXiv preprint arXiv:2505.05335*, 2025.

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

Yifei Xin, Zhihong Zhu, Xuxin Cheng, Xusheng Yang, and Yuexian Zou. Audio-text retrieval with transformer-based hierarchical alignment and disentangled cross-modal representation. *arXiv preprint arXiv:2409.09256*, 2024.

Ge Zhu, Jordan Darefsky, and Zhiyao Duan. Cacophony: An improved contrastive audio-text model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

A Related Work

Audio-Text Retrieval constructs a shared embedding space by contrastive learning, pulling matched audio-text pairs together and pushing unmatched pairs apart. Following CLIP (Radford et al., 2021), CLAP (Elizalde et al., 2023) has become the standard foundation, together with contemporaneous contrastive audio-text models (Wu et al., 2022; Guzhov et al., 2022). Subsequent Global methods refine this clip-level matching. CLAP-LAION (Wu et al., 2023) adds feature fusion and keyword enhancement for variable-length inputs. Cacophony (Zhu et al., 2024) augments retrieval with auxiliary objectives. SupCLAP (Luo et al., 2025) introduces Support Vector Regularization to stabilize the alignment trajectory. Ghosh et al. (2023) strengthen discrimination against hard negatives at the global level. These approaches remain coarse-grained: a single embedding represents the entire clip without separate representations for local acoustic events.

Fine-grained ATR instead decomposes auditory scenes into local units and aligns them with linguistic tokens. GPA (Xie et al., 2024) generates semantic prototypes from audio frames and words via dynamic pooling. THA-DCR (Xin et al., 2024) introduces hierarchical alignment and disentangled cross-modal factors. FLAM (Wu et al., 2025) applies frame-wise contrastive learning with logit adjustment. ASK (Fu et al., 2025) injects multi-grained out-of-batch knowledge to mitigate representation drift. Fang et al. (2025) convert audio prototypes into semantic identifiers for generative retrieval. These methods improve local interaction, yet two limitations persist.

Without geometric constraints, dynamically pooled prototypes are prone to Prototype Collapse (Jing et al., 2021), so fine-grained matching reverts to global alignment. Mini-batch contrastive training restricts negatives to the current batch, limiting the supply of high-similarity negatives for fine-grained discrimination (Kalantidis et al., 2020). Out-of-batch knowledge as used by ASK enlarges the negative pool but does not regularize prototype geometry. SARA addresses both limitations jointly through Stable Anchors, via Orthogonal Prototypes and Optimal Transport, and Rigorous Alignment, via Discriminative Hard Negative Mining from a momentum-updated queue.

B Roadmap of Appendix Experiments.

Appendix C provides the annotation procedure and distributions by split to complement the joint-distribution analysis in the main text. Appendix D examines the effect of γ on prototype geometry and retrieval, while Tab. 5 reports detailed results for queue size, mining budget, interpolation budget, OT, and prototype number. Appendix F reports memory usage and step time across configurations.

Appendix G presents attention distributions for GPA and GPA+HN. Appendix H examines semantic fragmentation at $K = 6$, and Appendix I examines the **Entity A + Relation + Entity B** specialization at $K = 3$. Appendix J further studies retrieval scores and prototype attention under entity substitution, changes in event order, and changes in caption richness. Finally, Appendix K discusses orthogonal regularization, geometric metrics, and prototype specialization.

C Event Density on AudioCaps

We use the DeepSeek V4 API to annotate every caption in the AudioCaps train, val, and test splits with acoustic events and the relations connecting them. Events describe sound-producing activities, while relations specify connections between these activities, such as temporal order or co-occurrence. For example, in "speech followed by crying," speech and crying denote two acoustic events, and followed by expresses their temporal relation, illustrating a two-event, one-relation structure. We count the event and relation labels to obtain the marginal distributions by split and the overall counts in Fig. 4. The corresponding joint distribution is presented in Fig. 2 in the main text.

Two events and one relation are the most frequent event count and relation count, respectively, alongside a substantial set of captions containing a single event and no relation. The test split contains a higher proportion of captions with three or more events than train and val. These distributions show the variation beyond the most frequent overall **2 Events + 1 Relation** combination and complement the prototype-number analysis in the main text.

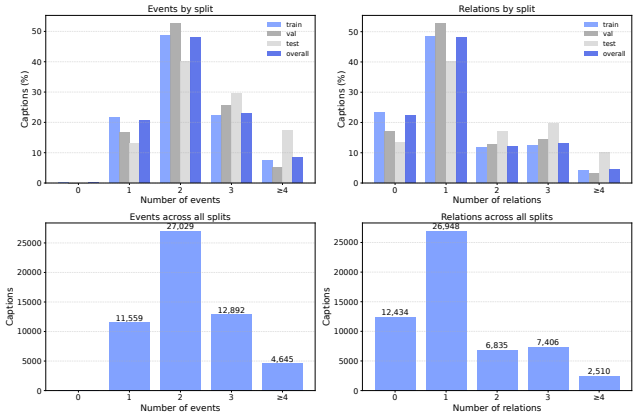


Figure 4: **Event and Relation Counts on AudioCaps.** Top: event-count and relation-count distributions for train, val, test, and overall. Bottom: overall event and relation counts.

D THE TRADE-OFF OF ORTHOGONALITY

We vary the orthogonal weight γ on AudioCaps and jointly examine retrieval performance and prototype geometry. Tab. 4 reports five settings from $\gamma = 0$ to $\gamma = 2.0$. As γ increases, APCS decreases and ρ increases in both modalities, while R@1 peaks at $\gamma = 1.0$ in both retrieval directions. These experiments characterize the relationship between orthogonal regularization strength, prototype diversity, and retrieval performance.

Table 4: Ablation study on the orthogonal weight γ on AudioCaps. We report retrieval performance (R@1) alongside geometric metrics (APCS and ρ) for both audio and text modalities.

Setting	**Retrieval Performance**		**Audio Geometry**		**Text Geometry**	
	A2T R@1	T2A R@1	APCS ↓	ρ ↑	APCS ↓	ρ ↑
$\gamma = 0.0$	52.94±0.75	41.47±0.23	1.00±0.01	0.15±0.02	1.00±0.00	0.03±0.01
$\gamma = 0.5$	52.98±1.37	41.58±0.43	0.95±0.01	0.37±0.03	0.99±0.01	0.13±0.02
$\gamma = 1.0$ **(Ours)**	**55.84±1.15**	**43.09±0.56**	0.21±0.01	0.96±0.01	0.17±0.02	0.98±0.01
$\gamma = 1.5$	52.87±0.55	42.61±0.80	0.06±0.02	1.00±0.00	0.05±0.01	1.00±0.00
$\gamma = 2.0$	53.05±0.61	42.57±0.28	0.05±0.02	1.00±0.00	0.04±0.02	1.00±0.00

Under-Regularization with $\gamma \leq 0.5$. At $\gamma = 0$, APCS reaches 1.00 in both modalities, with Audio and Text ρ of 0.15 and 0.03, indicating that prototypes concentrate along nearly the same direction. Increasing γ to 0.5 raises Audio and Text ρ to 0.37 and 0.13, while APCS remains high at 0.95 and 0.99. A2T R@1 changes only from 52.94% to 52.98%, and T2A R@1 from 41.47% to 41.58%. In both settings, highly correlated prototype directions and low effective rank indicate persistent *Prototype Collapse*.

Optimal Balance with $\gamma = 1.0$. Increasing γ to 1.0 reduces Audio and Text APCS to 0.21 and 0.17 and raises ρ to 0.96 and 0.98. At the same time, A2T and T2A R@1 increase to 55.84% and 43.09%, respectively, the highest values among the evaluated settings. We refer to this combination of low directional correlation and high effective rank as the **Quasi-Orthogonal State**. In this setting, restored geometric diversity accompanies improved retrieval in both directions.

Over-Regularization with $\gamma \geq 1.5$. At $\gamma = 1.5$, Audio and Text APCS decrease further to 0.06 and 0.05, and ρ reaches 1.00 in both modalities; A2T and T2A R@1 fall to 52.87% and 42.61%, respectively. At $\gamma = 2.0$, APCS decreases to 0.05 and 0.04, while R@1 remains below the $\gamma = 1.0$ setting at 53.05% and 42.57%. Relative to the retrieval peak, these two settings achieve stronger geometric separation but lower retrieval performance. We therefore adopt $\gamma = 1.0$ on AudioCaps, where orthogonal regularization maintains prototype diversity while achieving the highest R@1 in both directions.

Table 5: Detailed ablation results on AudioCaps. We report mean and standard deviation of R@1 and R@5 over three seeds. Full SARA uses $K = 3$, $M = 7200$, $N = 720$, and $S = 360$ and serves as the shared reference for each group.

Model Setting	A2T		T2A	
	R@1	R@5	R@1	R@5
Full SARA	55.84±1.15	84.36±0.31	43.09±0.56	76.82±0.46
$M = 0$	49.60±0.94	80.22±0.87	39.66±1.04	76.37±0.75
$M = 2400$	53.60±1.01	83.94±0.61	42.93±0.45	77.32±0.65
$M = 4800$	56.30±0.75	83.59±0.21	43.26±0.61	77.37±0.24
$N = 0$	54.30±1.73	83.56±1.00	42.86±0.79	77.02±0.81
$N = 360$	53.40±1.81	84.08±0.66	42.97±0.19	76.93±0.18
$S = 0$	52.53±1.33	82.83±0.64	42.81±0.45	76.77±0.33
$S = 720$	55.94±1.24	83.73±0.42	42.57±0.46	77.37±1.10
w/o OT	53.19±1.27	83.42±0.47	42.43±0.39	77.01±0.49
$K = 6$	53.29±1.03	84.71±1.05	43.26±0.58	77.56±0.67
$K = 9$	54.55±0.96	85.69±0.48	43.19±0.48	77.94±0.44

E Detailed Ablation Results

Impact of Momentum Queue Size. Tab. 5 compares retrieval results across queue sizes. Relative to $M = 0$, both $M = 4800$ and $M = 7200$ improve A2T R@1 by more than six percentage points and T2A R@1 by more than three percentage points. $M = 4800$ achieves higher mean R@1 in both directions and higher T2A R@5, whereas the default $M = 7200$ achieves higher A2T R@5. These results support the use of the queue-based pipeline and show that the effect of queue size differs across retrieval metrics.

Impact of Hard Negative Mining N. The N sweep fixes $S/N = 0.5$ when mining is enabled, so the interpolation budget increases together with the candidate budget. $N = 0$ retains random sampling from the queue and disables interpolation. $N = 720$ achieves A2T R@1 of 55.84%, exceeding 54.30% at $N = 0$ and 53.40% at $N = 360$, and yields the highest A2T R@5 and T2A R@1 among these three configurations. $N = 0$ instead achieves the highest T2A R@5. The default candidate budget therefore improves mean R@1, while the retrieval metrics do not improve uniformly with the mining budget.

Impact of Interpolation Budget S. With $N = 720$ fixed, increasing S from 0 to 360 raises A2T R@1 from 52.53% to 55.84% and R@5 from 82.83% to 84.36%, with improvements in both T2A metrics as well. Increasing S further to 720 adds only 0.10 percentage points to A2T R@1, reduces A2T R@5 and T2A R@1, and increases T2A R@5. We retain $S = 360$ as the default: it achieves similar A2T R@1 with half as many synthesized negatives, while maintaining higher A2T R@5 and T2A R@1.

Impact of Prototype Number K. Tab. 5 compares $K = 3$, 6, and 9. Among these settings, $K = 3$ achieves the highest A2T R@1 of 55.84%, compared with 53.29% and 54.55% at $K = 6$ and $K = 9$, respectively. Larger K improves R@5 in both directions, while T2A R@1 varies from 43.09% to 43.26%. We therefore retain $K = 3$ as a compact configuration that achieves the highest A2T R@1 among the evaluated prototype counts. In the $K = 6$ samples analyzed in Appendix H, additional prototypes exhibit repeated attention to the same content and overlapping representations of the same event. These observations characterize *Semantic Fragmentation*, which the appendix examines through audio and text attention visualizations.

F Efficiency Analysis of Momentum Queue and Hard Negative Mining

A recurring concern with momentum queue based methods and Top-N hard negative mining is whether the additional memory and compute cost would erode their practical value. To address

this concern, we benchmark SARA's queue-related computation on a single NVIDIA RTX 4090 with batch size $B = 24$, prototype number $K = 3$, embedding dimension $D = 1024$, queue size $M = 7200$, mining budget $N = 720$ and interpolation budget $S = 360$ as the default operating point, and sweep each of the five axes M, N, S, K and B around this default. The benchmark uses synthetic global and prototype representations and covers mining, interpolation, negative scoring, and backpropagation to query representations. It excludes encoder computation, momentum updates, and optimizer updates. We report the static queue footprint, the measured per-step peak GPU memory, the mining time and the benchmark step time averaged over 100 steps in Tab. 6.

Table 6: Efficiency of SARA's queue-related computation under characteristic settings for each swept axis. Mining measures one-direction queue scoring, Top-N selection, and interpolation. Step measures mining and negative scoring in both directions followed by backpropagation. Times are reported in milliseconds. Queue is the static memory of the momentum bank, and Peak is the maximum GPU footprint during the benchmark step. All measurements are on a single RTX 4090 with `float32`.

Sweep	Setting	B	K	M	N	S	Queue (MB)	Peak (MB)	Mining (ms)	Step (ms)
Default	B24 K3 M7200 N720 S360	24	3	7200	720	360	225.00	1774.28	4.68	18.44
M	M=1800	24	3	1800	720	360	56.25	1605.79	4.05	17.21
M	M=3600	24	3	3600	720	360	112.50	1662.20	4.26	17.66
M	M=14400	24	3	14400	720	360	450.00	2000.69	5.45	19.94
M	M=28800	24	3	28800	720	360	900.00	2454.00	7.95	24.94
N	N=180	24	3	7200	180	360	225.00	1129.79	3.13	10.81
N	N=360	24	3	7200	360	360	225.00	1332.33	3.65	13.38
N	N=1440	24	3	7200	1440	360	225.00	2792.27	6.62	28.16
N	N=2160	24	3	7200	2160	360	225.00	3810.10	8.55	37.79
S	S=0	24	3	7200	720	0	225.00	1262.20	2.13	10.34
S	S=180	24	3	7200	720	180	225.00	1516.99	3.70	14.98
S	S=720	24	3	7200	720	720	225.00	2417.72	6.56	25.21
S	S=1440	24	3	7200	720	1440	225.00	3774.75	10.11	38.03
K	K=1	24	1	7200	720	360	112.50	877.80	2.38	10.15
K	K=2	24	2	7200	720	360	168.75	1306.95	3.52	13.85
K	K=4	24	4	7200	720	360	281.25	2243.38	5.85	22.83
K	K=6	24	6	7200	720	360	393.75	3184.43	8.32	32.85
K	K=8	24	8	7200	720	360	506.25	4133.60	11.18	42.37
B	B=8	8	3	7200	720	360	225.00	754.93	1.90	6.87
B	B=16	16	3	7200	720	360	225.00	1261.60	3.27	12.66
B	B=48	48	3	7200	720	360	225.00	3304.18	8.81	35.70

Queue size M is cheap. The static queue memory scales strictly linearly with M, rising from 56.25 MB at $M = 1800$ through 225 MB at $M = 7200$ to 900 MB at $M = 28800$, in exact proportion to the queue length. The queue stores detached historical features, avoiding repeated encoding and gradient computation for queued representations while retaining the cost of queue scoring and candidate selection. The peak memory and step time grow much more slowly: peak moves from 1605.79 MB to 2454.00 MB and step time from 17.21 ms to 24.94 ms as M spans a 16× range, which corresponds to an average increase of approximately 0.62 ms per doubling of M at $M \leq 7200$ and about 3 ms per doubling beyond that, where the mining kernel starts to dominate. At our default $M = 7200$ the queue takes 225 MB, about 12.7% of the 1774 MB per-step peak, while the ablation in Tab. 5 shows a clear jump of more than six A2T R@1 points relative to $M = 0$. The ratio of accuracy gain to wall-clock overhead is therefore strongly favorable in this regime.

Mining budget N and interpolation budget S are the actual drivers. Increasing N and S enlarges the set of negative representations scored against each query, increasing matching and backpropagation costs. These negatives are drawn or interpolated from detached queue representations and require no additional encoding. The measurements in Tab. 6 show a nearly perfectly linear response on both axes. For N the peak grows from 1129.79 MB at $N = 180$ to 3810.10 MB at $N = 2160$, and the step time rises almost proportionally from 10.81 ms to 37.79 ms, giving a slope of roughly 1.37 ms per 100 extra negatives. For S the peak climbs from 1262.20 MB at $S = 0$ to 3774.75 MB at $S = 1440$ and the step time from 10.34 ms to 38.03 ms, again at a near-constant cost per synthesized sample. Our default choice of $N = 720$ at 10% of M and $S = 360$ at 50% of N sits at a soft knee of this curve: mining time at the default is 4.68 ms, compared with 2.13 ms when interpolation is removed and 10.11 ms when S is pushed to 1440. The ablations in Tab. 5 show that this moderate budget is already sufficient to realize the top-1 gains of hard negative mining and

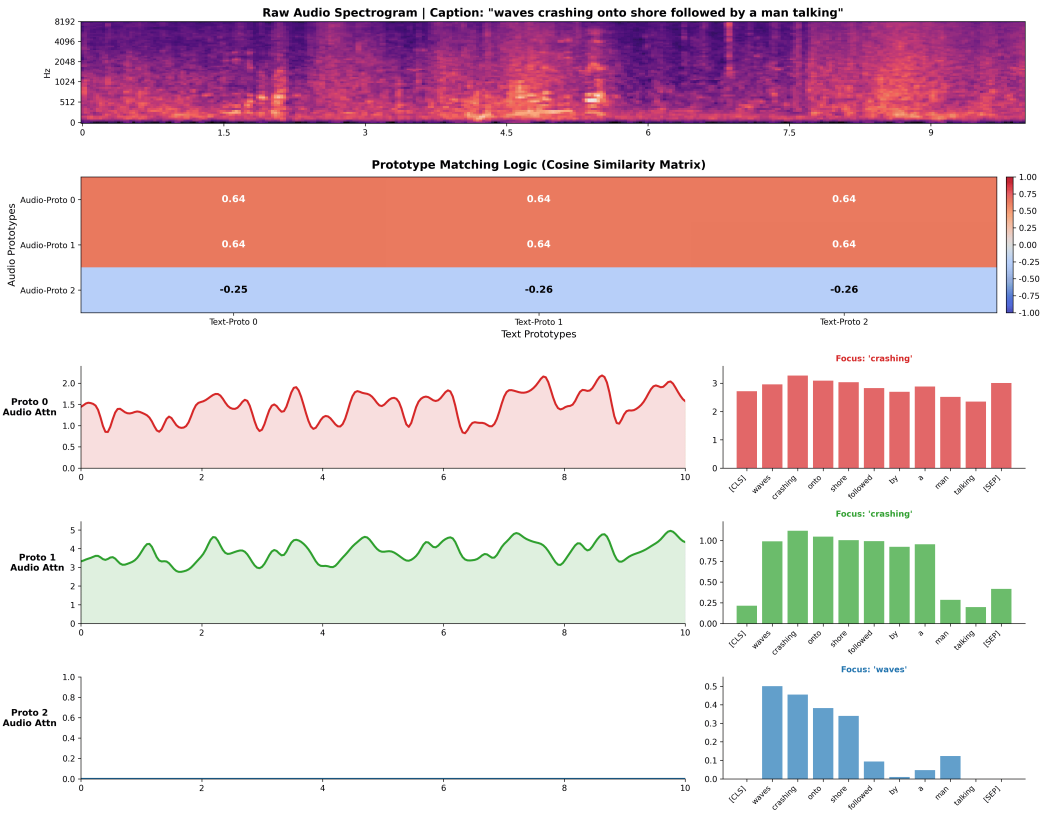


Figure 5: **Visual Diagnosis of the Baseline (GPA) Model.** Sample (a): "waves crashing onto shore followed by a man talking". Audio-Proto 2 has near-zero pooling weights across the entire temporal axis, while the remaining audio prototypes broadly cover the audio sequence. The text prototype weights overlap on wave-related tokens or broadly cover the caption, with neither modality exhibiting clear event decoupling.

manifold densification, so we see no empirical motivation to spend the extra compute of larger N or S.

Prototype count K and batch size B dominate the peak. The step-peak memory is close to linear in both K and B over the tested range. Increasing these parameters enlarges the representation tensors and the pairwise prototype similarity matrices used for scoring. Sweeping K from 1 through 2, 3, 4, 6 to 8 grows the peak from 877.80 MB to 4133.60 MB, with each additional prototype adding roughly 460 MB and 4.60 ms of step time. Sweeping B from 8 through 16, 24 to 48 grows the peak from 754.93 MB to 3304.18 MB and the step time from 6.87 ms to 35.70 ms, again in near-linear proportion to B. These increases reflect the cost of storing and matching prototype representations and backpropagating through the query-side scoring computation. Taken together, the measurements show that expanding the queue incurs moderate additional cost over the tested range, while the negative budgets, prototype count, and batch size have substantial effects on the measured memory and runtime.

G Visual Diagnosis of Prototype Collapse

This section extends the diagnosis in Sec. 4.3 through visual analysis of audio and text prototypes. We compare GPA and GPA+HN on two samples containing compound events, examining whether the pooling weights of $K = 3$ prototypes establish a clear division between events. We further analyze *Prototype Collapse* using the geometric metrics in Tab. 2.



Figure 5: **Visual Diagnosis of the Baseline (GPA) Model.** Sample (b): "a young woman speaks loudly with a toilet flushing sound". The audio-side Partial Failure in Fig. 5(a) recurs, with near-zero pooling weights for Audio-Proto 2 and broad temporal coverage by the remaining prototypes. The text weights broadly cover the caption or emphasize tokens from both events, failing to establish a clear division between events.

Visualization of GPA: Partial Failure and Lack of Event Decoupling. Figs. 5(a) and 5(b) show audio-side Partial Failure in both samples. Audio-Proto 2 has near-zero pooling weights across the entire temporal axis, exhibiting a Dead Prototype at the input aggregation stage. Audio-Proto 0 and Audio-Proto 1 maintain broad coverage over the audio sequence without a clear division between events. The audio-side failure therefore comprises both prototype inactivity and a lack of event decoupling among the active prototypes.

The text side also lacks a clear division between events. In Fig. 5(a), Text-Proto 1 and Text-Proto 2 both emphasize wave-related tokens, while Text-Proto 0 broadly covers the caption. In Fig. 5(b), Text-Proto 1 emphasizes flushing-related tokens, Text-Proto 2 emphasizes selected tokens from both events, and Text-Proto 0 maintains broad coverage. These weight distributions show overlapping event attention and fail to clearly decompose the compound captions across prototypes.

These visualizations complement the geometric diagnosis in Tab. 2 at the input aggregation level. GPA has Text APCS of 1.00 and ρ of 0.07, indicating that the final text prototypes have nearly identical directions. Audio APCS and ρ are 0.61 and 0.43, respectively, while both audio samples exhibit an inactive prototype alongside active prototypes with broad input coverage. Together, these results reveal partial failure of audio prototypes, a lack of clear event decoupling in both modalities, and geometric collapse among the final text prototypes.

Visualization of GPA+HN: Active Prototypes Still Lack Event Decoupling. In Figs. 6(a) and 6(b), all three audio prototypes have nonzero pooling weights. Their temporal coverage overlaps substantially, and Audio-Proto 0 and Audio-Proto 2 exhibit similar temporal patterns in both sam-

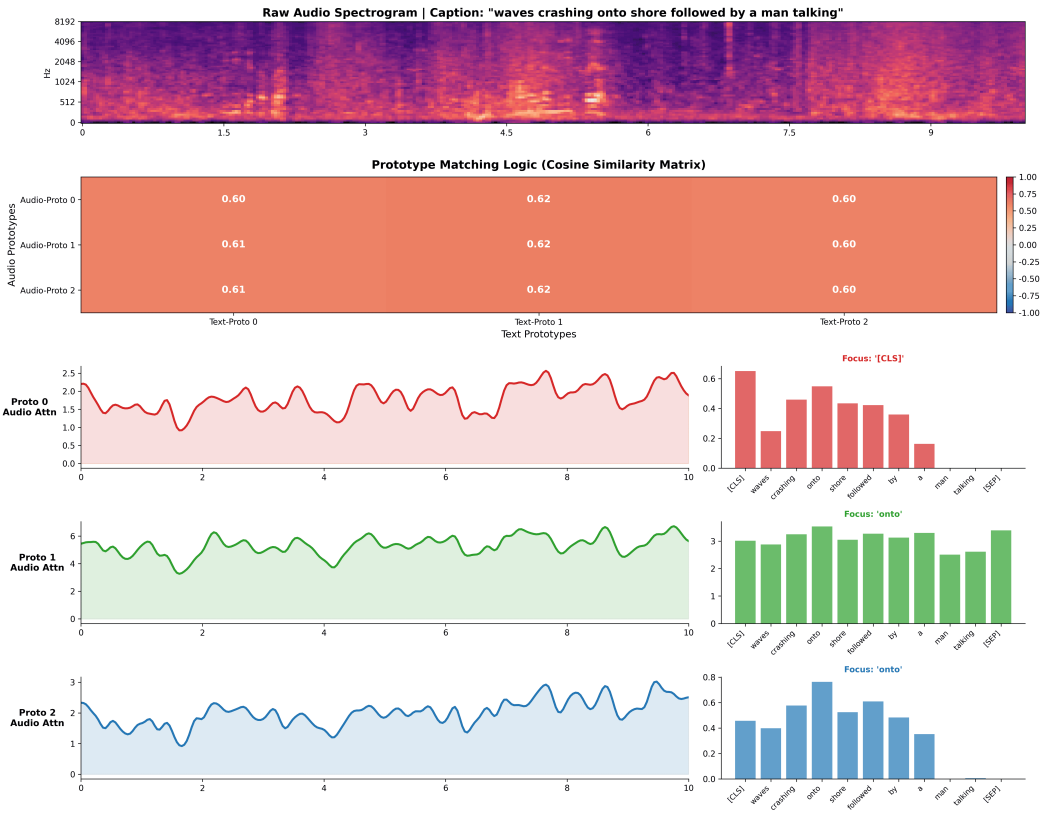


Figure 6: **Visual Diagnosis of Baseline + HN.** Sample (a): "waves crashing onto shore followed by a man talking". All three audio prototypes have nonzero pooling weights but highly overlapping temporal coverage, with similar temporal patterns for Audio-Proto 0 and Audio-Proto 2. The text prototypes emphasize tokens in the first part of the caption or broadly cover the caption, with both modalities lacking clear event decoupling.

ples. Although all three prototypes participate in input aggregation, they still fail to establish a clear division between acoustic events.

The text side retains the same decoupling problem. In Fig. 6(a), Text-Proto 0 and Text-Proto 2 primarily emphasize tokens in the first part of the caption, while Text-Proto 1 broadly covers the caption. In Fig. 6(b), Text-Proto 0 and Text-Proto 2 assign higher weights to *loudly* and selected flushing-related tokens, while Text-Proto 1 broadly covers both events. These distributions differ in their local weights, but their event attention remains overlapping and fails to establish clear event decoupling.

The geometry of the final prototypes further reveals collapse. In Tab. 2, GPA+HN has Audio and Text APCS of 1.00, with ρ of 0.12 and 0.07, respectively, indicating highly aligned prototype directions in both modalities. Meanwhile, A2T R@1 increases from 49.53% to 52.14% and T2A R@1 from 40.04% to 43.09% relative to GPA. Hard Negative Mining therefore improves retrieval performance, while the geometric diagnosis and sample visualizations show persistent *Prototype Collapse* and a lack of clear event decoupling, respectively. This supports the central finding in Sec. 4.3: prototype diversity and discriminative retrieval training need to be addressed jointly.

H Semantic Fragmentation with Six Prototypes

Tab. 5 shows that increasing K from 3 to 6 reduces A2T R@1 from 55.84% to 53.29%, while T2A R@1 increases from 43.09% to 43.26% and R@5 improves in both directions. To examine how additional prototypes distribute attention over the inputs, we analyze the audio and text weights

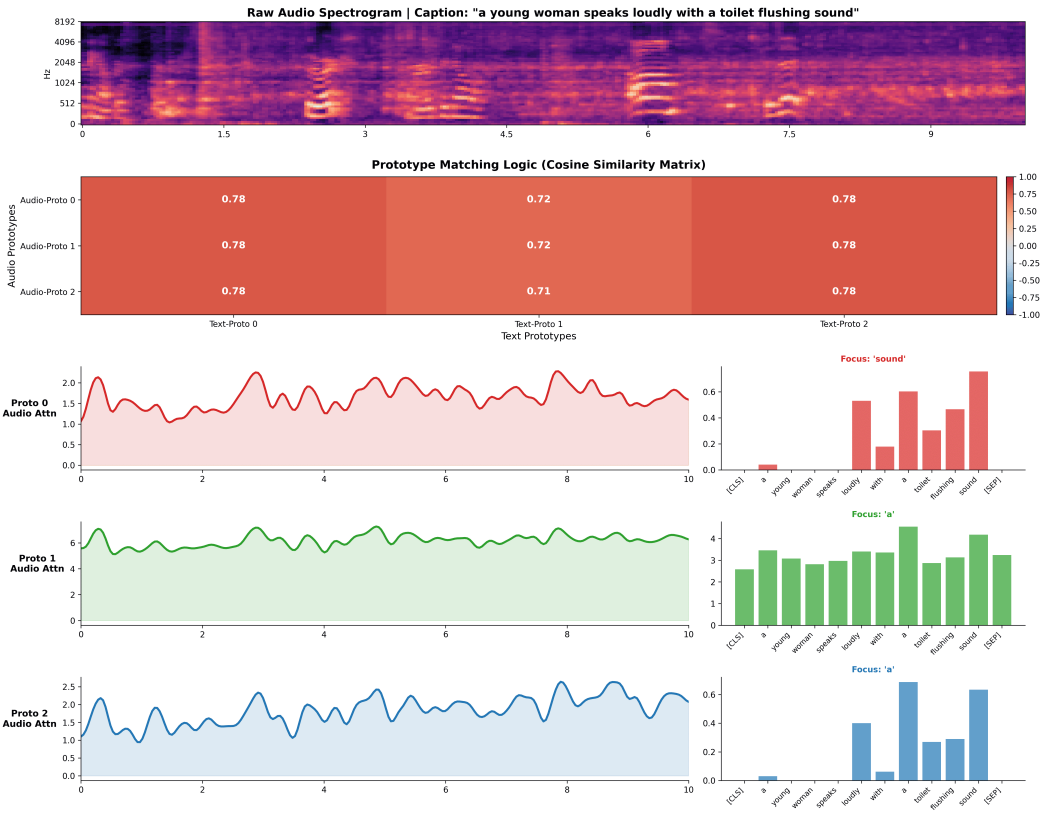


Figure 6: **Visual Diagnosis of Baseline + HN.** Sample (b): "a young woman speaks loudly with a toilet flushing sound". The audio prototypes broadly cover the sequence, with highly similar temporal patterns among some prototypes. The text prototype weights overlap on *loudly* and flushing-related tokens or broadly cover both events, showing that the active prototypes still lack a clear division between events.

of two compound-event samples at $K = 6$. The $K = 6$ cases exhibit *Semantic Fragmentation*: attention to the same event is distributed across overlapping prototypes, while multiple prototypes repeatedly emphasize the same tokens.

Repeated Token Attention and Fragmentation of Concepts. In Fig. 7(a), Text-Proto 0, 1, and 3 assign their highest weights to [SEP], while Text-Proto 2 concentrates its weights on *onto*. Meanwhile, Text-Proto 4 and Text-Proto 5 both attend to *man* and *talking*, with different relative weights. The same speech event is thus represented by two prototypes with overlapping attention. In Fig. 7(b), Text-Proto 0, 1, 3, and 4 assign their highest weights to [SEP], while Text-Proto 0, 1, and 3 also emphasize flushing-related tokens. These distributions show that additional prototypes are allocated to repeated token attention and overlapping event representations.

Attention Overlap on the Audio Side. Audio-Proto 2 and Audio-Proto 3 exhibit highly similar temporal weight patterns in both samples. In Fig. 7(a), they peak at several shared temporal locations. In Fig. 7(b), their dominant peaks also occur at the same locations. This repeated temporal attention reveals prototype redundancy at the audio aggregation stage.

Semantic Fragmentation and Prototype Redundancy. Both $K = 6$ samples exhibit repeated token attention, overlapping representations of the same event across prototypes, and repeated temporal attention on the audio side. These findings show that the additional prototypes fail to establish a correspondingly richer division between events in the analyzed samples, instead exhibiting *Semantic Fragmentation* and prototype redundancy.

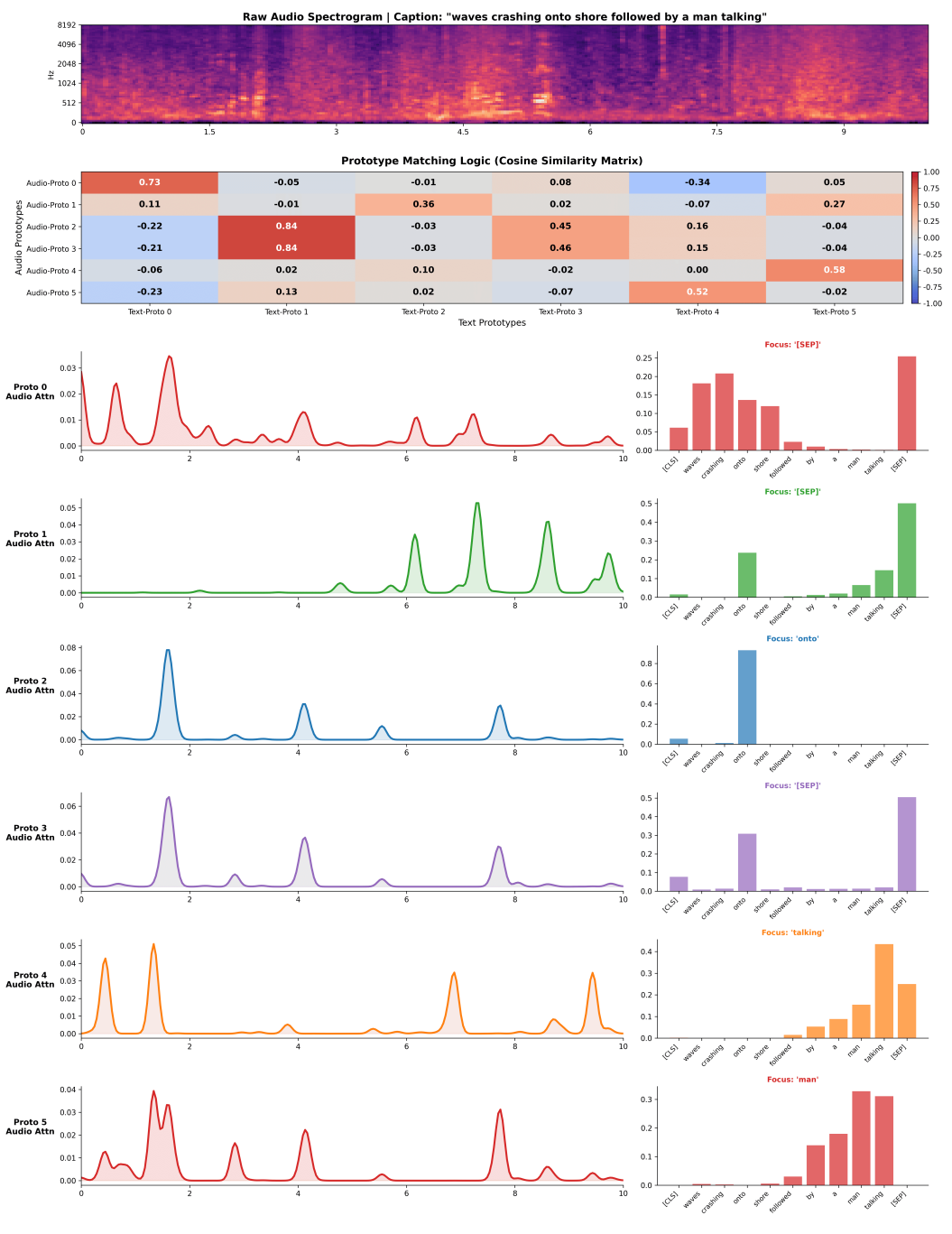


Figure 7: **Illustration of Semantic Fragmentation (K = 6).** Sample (a): ”waves crashing onto shore followed by a man talking”. Text-Proto 0, 1, and 3 assign their highest weights to [SEP], while Text-Proto 2 concentrates on *onto*. Text-Proto 4 and Text-Proto 5 have overlapping attention to *man talking*, and Audio-Proto 2 and Audio-Proto 3 exhibit highly similar temporal weight patterns, revealing Semantic Fragmentation and prototype redundancy in both modalities.

I Qualitative Analysis: Prototypes as Semantic Bottlenecks

We examine the same two compound-event samples used in Appendices G and H to study prototype specialization in SARA at $K = 3$. Figs. 8(a) and 8(b) show the audio and text attention weights from the prototype construction modules. We use p_1, p_2, and p_3 to denote Text-Proto 0, Text-Proto

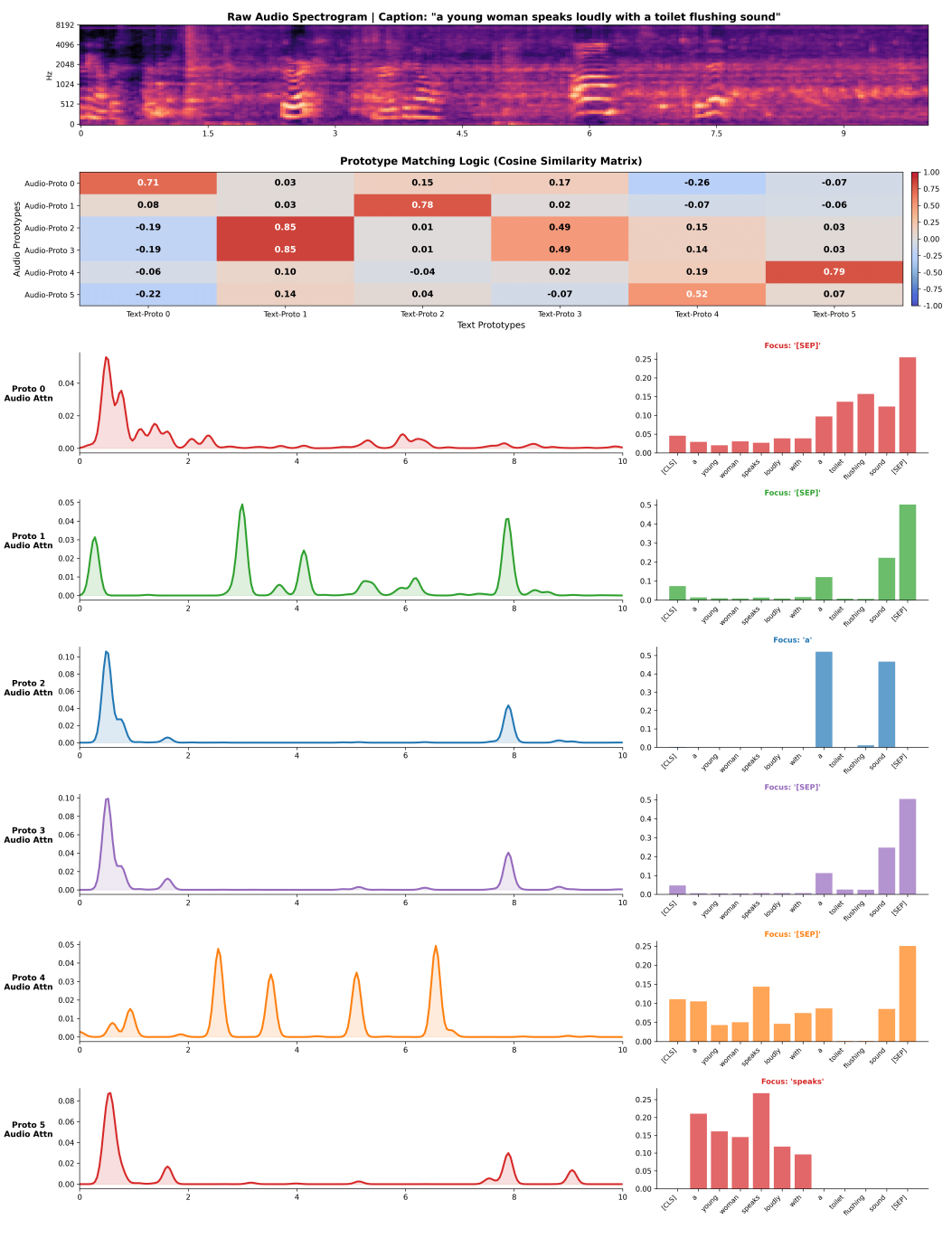


Figure 7: **Illustration of Semantic Fragmentation (K = 6).** Sample (b): "a young woman speaks loudly with a toilet flushing sound". Text-Proto 0, 1, 3, and 4 assign their highest weights to [SEP], while multiple prototypes also attend to flushing-related tokens. Audio-Proto 2 and Audio-Proto 3 exhibit highly similar temporal weight patterns, showing repeated attention across both samples together with Fig. 7(a).

1, and Text-Proto 2, respectively. The text weights reveal a clear **Entity A + Relation + Entity B** decomposition, while the audio weights show selective attention to different temporal locations.

Emergent Functional Specialization. SARA uses no prototype-role labels during training. In these two samples, the three text prototypes form a clear division: two prototypes attend to different

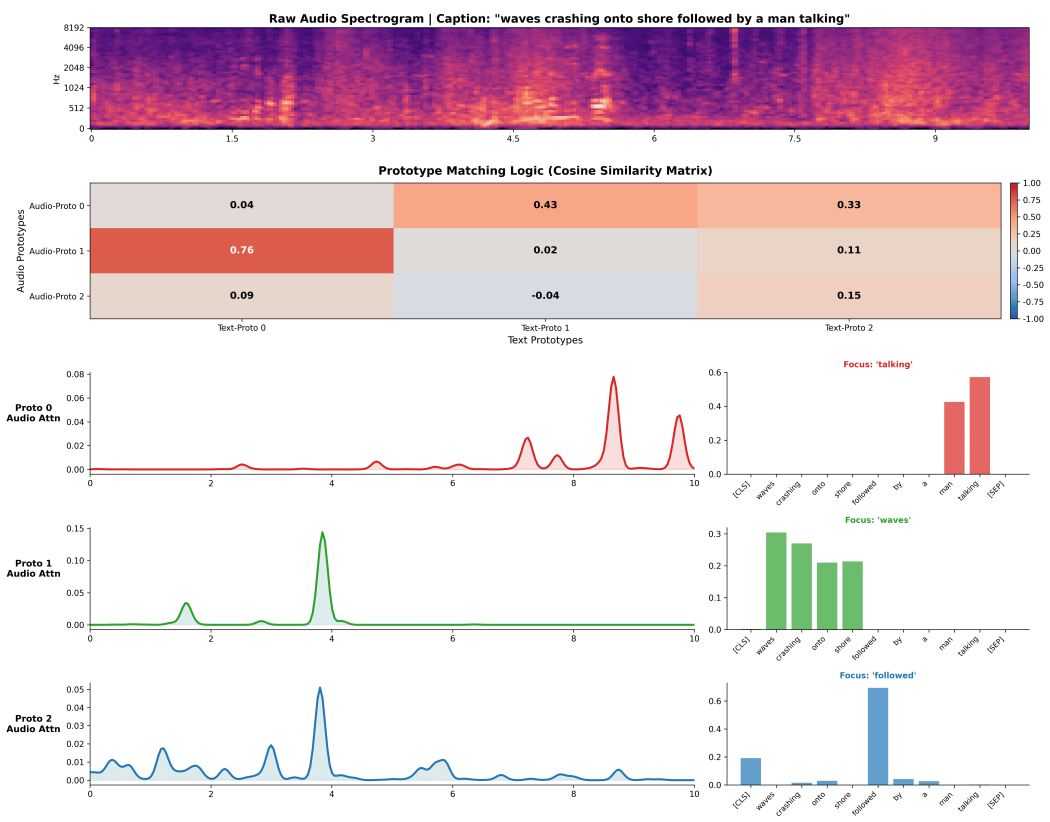


Figure 8: **Interpretability of SARA ($K = 3$).** Sample (a): "waves crashing onto shore followed by a man talking". Text prototypes p_1 and p_2 attend to *man talking* and *waves crashing onto shore*, respectively, while p_3 attends to the temporal connective *followed*. This event and relation decomposition exhibits Emergent Functional Specialization; on the audio side, Audio-Proto 0 and Audio-Proto 1 concentrate their main weights in the later and earlier parts of the sequence, respectively.

events, and the third attends to the word connecting them. This division emerges during retrieval training and exhibits **Emergent Functional Specialization**.

Event-Specific Prototypes. In Fig. 8(a), p_1 concentrates on *man talking*, while p_2 attends to *waves crashing onto shore*, corresponding to the two acoustic events in the caption. In Fig. 8(b), p_1 attends to *young woman speaks*, and p_2 attends to *toilet flushing*. Both samples exhibit clear event decoupling, with related tokens within each event jointly attended to by a single prototype.

Relational Prototype as a Syntactic Anchor. In both samples, p_3 attends to the relation between events. In Fig. 8(a), it assigns its highest weight to *followed*, corresponding to the temporal order of the two events. In Fig. 8(b), it concentrates on *with*, corresponding to their co-occurrence. By attending to the word connecting the two events, this prototype serves as a **Syntactic Anchor** and completes the event and relation decomposition alongside the two Event-Specific Prototypes.

Temporal Attention on the Audio Side. The audio prototypes also exhibit distinct temporal weight patterns. In Fig. 8(a), Audio-Proto 0 has its dominant peaks in the later part of the sequence, Audio-Proto 1 has its dominant peaks in the earlier part, and Audio-Proto 2 covers multiple temporal locations. In Fig. 8(b), Audio-Proto 0 and Audio-Proto 1 have dominant peaks at different temporal locations, while Audio-Proto 2 has more distributed temporal coverage. These distributions show selective audio aggregation and, together with the clear semantic division on the text side, characterize the fine-grained representations learned by SARA.

Prototypes as Semantic Bottlenecks. In these cases, three prototypes organize each compound caption into two events and their relation, forming a compact semantic representation. The **Semantic Bottleneck** is expressed through a clear division within a limited prototype set: the Event-Specific

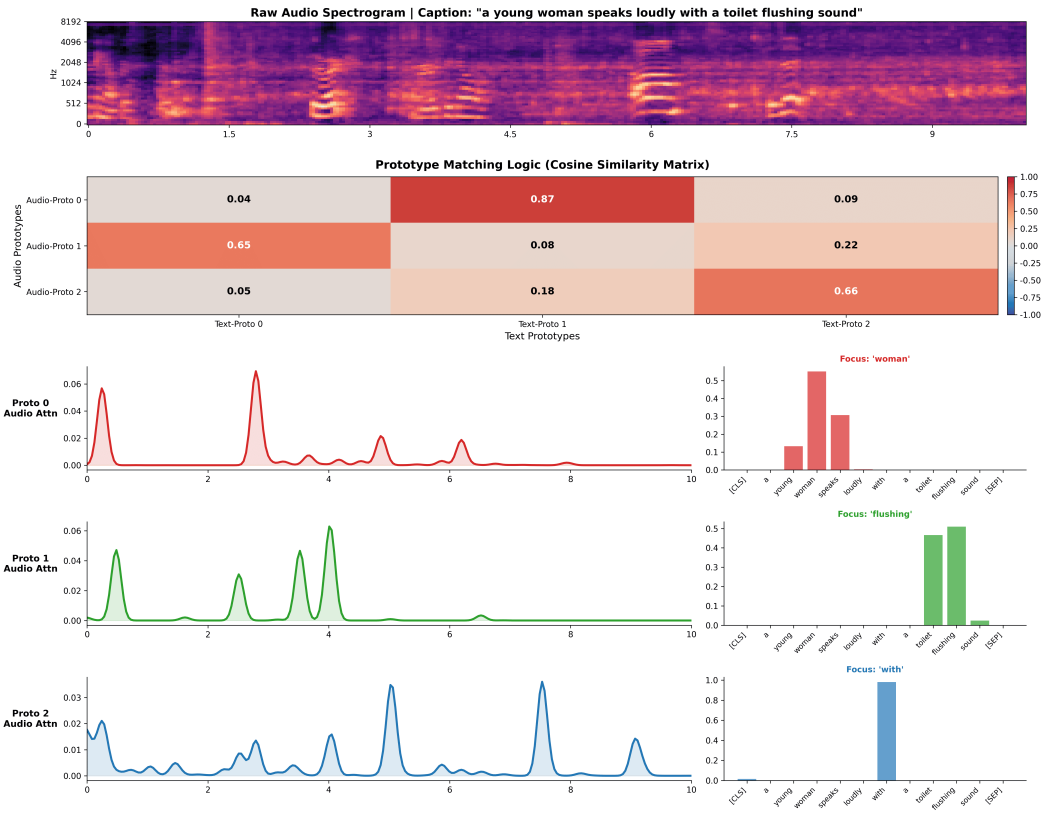


Figure 8: **Interpretability of SARA ($K = 3$).** Sample (b): "a young woman speaks loudly with a toilet flushing sound". Text prototype p_1 attends to *young woman speaks*, p_2 attends to *toilet flushing*, and p_3 attends to the co-occurrence connective *with*. This distribution again forms an Entity A + Relation + Entity B decomposition, while the audio prototypes exhibit distinct temporal attention patterns.

Prototypes aggregate content associated with each event, while the Relational Prototype attends to the connection between them.

Discussion on Event Density. Fig. 2 shows the joint distribution of event and relation counts in AudioCaps captions, with statistics by split provided in Appendix C. The most frequent combination is two events and one relation, occurring in 26,672 captions and corresponding to the **Entity A + Relation + Entity B** decomposition in the cases above. The dataset also contains captions with a single event and no relation, which have a simpler semantic structure, while captions with more events require representations of additional event content and their connections. A fixed set of three prototypes therefore faces different semantic allocation requirements across captions. The $K = 6$ cases in Appendix H further exhibit repeated attention and *Semantic Fragmentation*. Together with the retrieval results in Tab. 5, $K = 3$ provides a compact configuration for prototype allocation and semantic specialization while achieving the highest A2T R@1 among the evaluated prototype counts.

J Fine-grained Hard-Negative Discrimination via A + Relation + B Prototypes

To further analyze the **Entity A + Relation + Entity B** specialization in Appendix I, we construct hard negatives around two AudioCaps captions. The two samples cover event co-occurrence and temporal order: a woman speaking with a toilet flushing sound, and *speech followed by crying*. The former examines prototype responses to changes in speaker identity, acoustic attributes, and event

content. The latter further examines changes in matching when event content is retained but event order is reversed.

For each sample, we construct acoustic and richness hard negatives. Acoustic negatives change speaker identity, acoustic attributes, or event content, and reverse the event order in the temporal sample. Richness negatives remove or add events. This design examines how event prototypes respond to content changes, how event order affects cross-modal matching, and how prototypes are allocated when the event count changes. We jointly analyze caption rankings, OT-weighted prototype matching contributions, and text prototype attention to connect semantic specialization with retrieval discrimination.

The total score in the figures is the sum of the CLS and prototype scores. All score changes below are computed as the hard-negative score minus the ground-truth score. We use p_1, p_2, and p_3 for Proto 0, Proto 1, and Proto 2 in the figures, respectively.

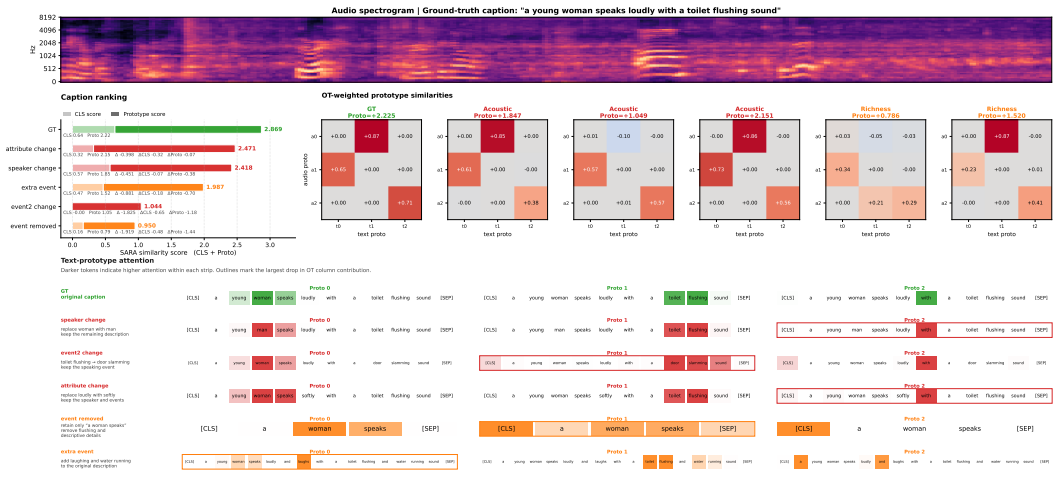


Figure 9: **Hard negative analysis.** Sample (a): *"a young woman speaks loudly with a toilet flushing sound"*. The bar chart shows total scores and their CLS and prototype components, the matrices show OT-weighted prototype matching contributions, and the lower heatmaps show token attention for the three text prototypes. The ground-truth caption receives a higher total score than every displayed hard negative, with complementary discrimination from the CLS and prototype components under acoustic-attribute and speaker-identity changes.

Event and Relation Specialization under Content Perturbations. For the ground-truth woman-and-flushing caption, p_1 attends to *young woman speaks*, p_2 to *toilet flushing*, and p_3 to *with*. For the temporal caption, the three prototypes attend to *speech*, *crying*, and *followed*, respectively. Both examples exhibit a clear event and relation division. The hard negatives further examine how this division responds to caption changes: entity substitution changes the content attended to by the corresponding event prototype, order reversal changes the positions of events in the sentence, and event addition or removal changes the semantic content allocated to the prototypes.

Entity Substitution and Prototype Migration. Replacing *woman* with *man* shifts the main attention of p_1 to *man*, while replacing *toilet flushing* with *door slamming* shifts p_2 to the substituted event tokens. In the temporal sample, replacing *speech* with *singing* and *crying* with *knocking* similarly shifts the attention of p_1 and p_2, respectively. The other event prototype continues to attend to the retained event, while p_3 maintains its main attention on the connective.

This correspondence reveals the role of prototype specialization: the same event prototype accommodates the substituted content while the remaining prototypes preserve their event and relation roles. Semantic selection is thereby connected to matching discrimination. The substituted event is still aggregated by the corresponding prototype, but its matching score against the original audio decreases. In the temporal sample, replacing *speech* with *singing* and *crying* with *knocking* decreases the prototype score by 0.58 and 0.91, respectively, showing that event prototypes follow new content while identifying its mismatch with the audio.

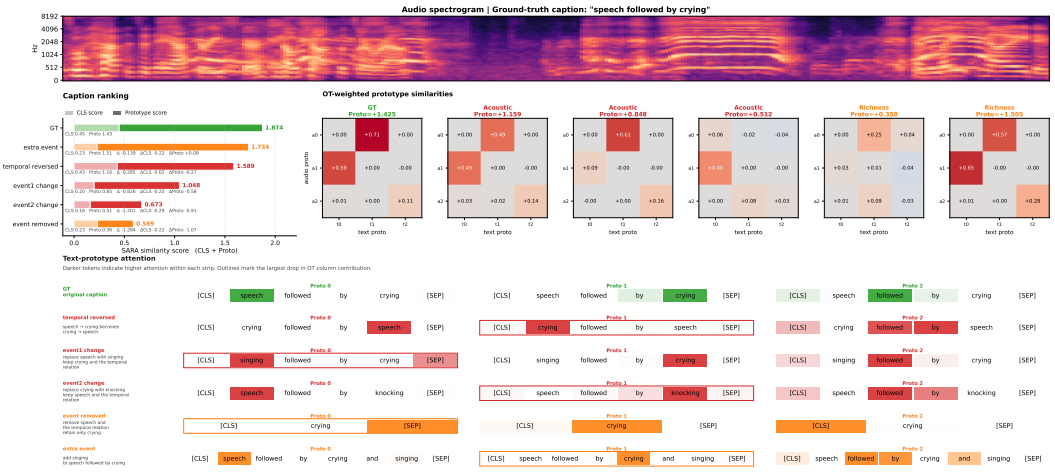


Figure 9: **Hard negative analysis.** Sample (b): *"speech followed by crying"*. Same layout as Fig. 9(a). After event-order reversal, p_1 and p_2 continue to attend to *speech* and *crying*, respectively, while the prototype score decreases by 0.27 and the CLS score by 0.02. Adding *singing* increases the prototype score by 0.08 but decreases the CLS score by 0.22, keeping the total score below the ground truth and demonstrating complementary contributions from the two components.

Event-Order Reversal and Content-Grounded Prototypes. Changing *speech followed by crying* to *crying followed by speech* exchanges the positions of the two event tokens, while p_1 continues to attend to *speech* and p_2 to *crying*. This directly demonstrates **content-grounded anchors**: through this order reversal, prototype event identity follows content rather than token position.

Although the event tokens are retained, the prototype score decreases by 0.27, the CLS score by 0.02, and the total score by 0.285. The two dominant event matching contributions in the OT-weighted matrix decrease from approximately 0.71 and 0.59 to 0.49 and 0.49, respectively, showing that the order change is reflected in cross-modal matching of the event prototypes. Together with the attention maps, this result reveals that content selection and relation-sensitive matching coexist: prototypes continue to aggregate tokens of the same event while distinguishing its match to the audio under different event orders. The prototype component thus provides the main score separation in this reversal case.

Complementary CLS and Prototype Scores. In Fig. 9(a), changing *loudly* to *softly* decreases the CLS score by 0.32 and the prototype score by 0.07. Changing *woman* to *man* decreases the CLS score by 0.07 and the prototype score by 0.38. The two components therefore provide greater score separation for the attribute and identity changes, respectively. In Fig. 9(b), adding *singing* to the ground-truth caption increases the prototype score from 1.425 to 1.505, while the CLS score decreases by 0.22 and the total score decreases by 0.139. The combined score ranks the ground-truth caption first in both candidate sets.

Event Count and Prototype Allocation. Reducing the temporal caption to the single event *crying* preserves the attention of p_2 to *crying*, while p_1 and p_3 shift to [SEP] and [CLS], respectively. The prototype score decreases by 1.07. Reducing the woman-and-flushing caption to *a woman speaks* preserves the attention of p_1 to the speech event, distributes the weights of p_2 over the remaining tokens, and shifts p_3 to [CLS], with a prototype-score decrease of 1.44. Adding events also redistributes attention: in the woman-and-flushing sample, the added *laughs* and *water running* attract attention from the two event prototypes, respectively. These cases demonstrate changes in prototype allocation with event density. For single-event captions, the original two-event and one-relation division reduces to attention over the retained event and special tokens; with additional events, multiple event contents share the existing prototypes. This observation complements the event-density discussion in Appendix I.

K DISCUSSION

Orthogonal Regularization. SARA controls inter-prototype correlations through the orthogonal loss weighted by γ. The AudioCaps sweep in Appendix D shows that $\gamma = 1.0$ achieves the highest R@1 in both directions among the evaluated settings, with Audio and Text APCS of 0.21 and 0.17. Increasing γ to 1.5 or 2.0 further reduces APCS but lowers R@1. Orthogonal regularization therefore serves to suppress Prototype Collapse, while retrieval performance depends on joint optimization of the orthogonal and contrastive objectives.

Geometric Metrics and Semantic Specialization. APCS and ρ characterize average directional similarity and effective rank among prototypes, respectively. Appendix G combines these metrics with pooling weights to reveal inactive audio prototypes in GPA and a lack of clear event decoupling in GPA and GPA+HN. Appendix H further shows repeated token attention and temporal audio attention at $K = 6$. Through entity substitution and event-order reversal, Appendix J connects prototype content selection to cross-modal matching scores. Together, these analyses establish a diagnosis spanning prototype geometry, input aggregation, and retrieval discrimination, making semantic specialization a model property examined through concrete behavior.

Prototype Specialization and Event Density. SARA learns prototype specialization through retrieval training without prototype-role labels. The compound-event cases in Appendix I exhibit a clear **Entity A + Relation + Entity B** structure. AudioCaps contains both single-event and multi-event captions, placing different content allocation demands on a fixed prototype set. The event-removal and event-addition experiments in Appendix J illustrate this variation. Some prototypes shift to special tokens when a caption is reduced to a single event, while additional event content shares the existing prototypes when events are added. These observations provide a concrete motivation for **Adaptive Prototype Activation** based on input content. Future work can investigate how to adjust the number of active prototypes according to event and relation content, adapting prototype allocation to different event densities.