SAYAKA: SEMANTIC ATTENTION YIELDS AUDIO–CONCEPT ALIGNMENT FOR FINE-GRAINED AUDIO–TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained audio–text retrieval requires models to precisely associate events described in captions with their corresponding acoustic evidence. Existing fine-grained alignment methods follow two paths. One introduces explicit frame-level supervision during training, relying on fixed event taxonomies that do not fully adapt to retrieval scenarios with open-vocabulary captions. The other learns implicit alignment between events and multiple fine-grained prototypes, but requires computing similarity scores with these prototypes at retrieval time, incurring noticeable overhead compared to the single-vector inner product of standard global audio–text retrieval. We propose SAYAKA, a fine-grained audio–text retrieval method that injects concept-level alignment structure only during training and fully reduces to a standard dual encoder at inference. SAYAKA parses captions into semantic concepts via lightweight text parsing; Semantically Conditioned Audio uses each concept as a query to produce conditioned audio tokens from a shared audio-token sequence through semantic attention. Soft-Positive Audio–Concept Alignment treats other concepts within the same caption as weak positives rather than negatives, while a concept separation constraint encourages distinct concepts to capture non-overlapping acoustic evidence. An adaptive audio recomposition module fuses multiple conditioned representations into a single vector aligned with the global audio embedding, enabling fine-grained retrieval using only a single vector at inference. We further construct momentum queues at both the sample and concept levels to continuously strengthen the model's ability to distinguish fine-grained conceptual differences across samples. Case studies and experimental results demonstrate that SAYAKA achieves fine-grained concept–audio alignment and surpasses all current audio–text retrieval methods on AudioCaps and Clotho in terms of R@1, while maintaining the retrieval efficiency of global audio–text retrieval.

1 INTRODUCTION

Audio–text retrieval (ATR) achieves bidirectional retrieval between language and audio through cross-modal representation alignment. Mainstream methods adopt dual encoders that independently encode audio and text into single global vectors and rank candidates by their inner product (Elizalde et al., 2023; Wu et al., 2023). Recent work has gradually shifted from coarse-grained alignment between global embeddings toward fine-grained alignment (Wu et al., 2025), decomposing acoustic scenes into discrete events and associating them with specific audio frames to achieve better interpretability and retrieval precision.

Existing fine-grained alignment work follows two paths, each introducing additional costs. The first introduces frame-level supervision on the audio side. FineLAP maps phrases to semantic clusters initialized from a fixed event ontology and constructs frame-level positive and negative labels through cluster-based sampling (Li et al., 2026b). FLAM programmatically synthesizes mixed audio and automatically generates open-vocabulary frame-level labels using synthesis boundaries (Wu et al., 2025). The frame-level supervision in these methods comes from additional annotations or data synthesis pipelines beyond paired captions. The former is constrained by a fixed event taxonomy, while the latter relies on large-scale audio synthesis rather than directly deriving supervision

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.

from naturally occurring audio–caption pairs in retrieval scenarios. The second retains local or prototype matching in retrieval scoring, such as text-conditioned attention pooling, multi-layer token interaction, or prototype similarity matrices (Mei et al., 2022; Xin et al., 2024; Xie et al., 2024). These scores depend on local interactions between the current query and candidate, introducing noticeable additional retrieval overhead compared to the single-vector inner product of global ATR.

We propose SAYAKA, a fine-grained audio–text retrieval method that injects concept-level alignment structure only during training and fully reduces to a standard dual encoder at inference. SAYAKA parses captions into a set of semantic concepts using a lightweight rule-based text parser, relying solely on paired captions without requiring external event ontologies or data synthesis pipelines. In Semantically Conditioned Audio, each concept's text features serve as queries to extract conditioned audio tokens from a shared audio-token sequence through semantic attention, providing concept-specific acoustic evidence. Concepts within the same caption are semantically related but point to different acoustic events; therefore, Soft-Positive Audio–Concept Alignment treats other concepts from the same caption as weak positives rather than negatives, while the concept separation constraint encourages distinct concepts to capture non-overlapping acoustic regions. An adaptive audio recomposition module fuses multiple conditioned representations into a single vector aligned with the global audio embedding, allowing concept-level training signals to directly shape the global representation used at inference, enabling fine-grained retrieval using only a single-vector inner product at inference time. We further construct momentum queues at both the sample and concept levels to mine hard negatives, continuously strengthening the model's ability to distinguish fine-grained differences among global and concept-level hard negatives across samples.

The contributions of this paper are summarized as follows:

1. We propose SAYAKA, which uses semantic concepts parsed from paired captions as queries to produce conditioned audio tokens through semantic attention and fuses them into a single vector aligned with the global audio embedding via adaptive recomposition. This allows fine-grained concept-level supervision to directly act on the global representation used at inference, enabling the model to benefit from fine-grained training without retaining the concept branch during retrieval.
2. We design Soft-Positive Audio–Concept Alignment and a concept separation constraint to model weak-positive and mutually exclusive relationships among concepts within the same caption, and construct momentum queues at both the sample and concept levels to continuously strengthen the model's ability to distinguish subtle conceptual differences across samples.
3. Case studies demonstrate the correspondence between concepts and acoustic evidence, while component and queue ablations validate the contribution of each design choice. Under a unified backbone, SAYAKA achieves the highest reported mean R@1 on all four retrieval directions of AudioCaps and Clotho, with the significant efficiency advantage of computing similarity scores through a single matrix inner product at retrieval time.

2 Related Work

CLAP (Elizalde et al., 2023) learns shared audio–text representations through contrastive pretraining, while LAION-CLAP (Wu et al., 2023) incorporates feature fusion for variable-length audio and keyword-to-caption augmentation. Cacophony (Zhu et al., 2024) combines masked audio pretraining with contrastive and auxiliary captioning objectives, while WavCaps (Mei et al., 2024) provides large-scale weakly labelled audio–caption pairs. WavText5K (Deshmukh et al., 2023) further examines how web audio–text training data affects retrieval. At the optimization level, SupCLAP (Luo et al., 2026) uses support vector regularization to control trajectory drift. Alongside these training improvements, research examines how audio–text models handle complex semantics and varied queries: DTU (Oncescu et al., 2024) and CompA (Ghosh et al., 2024) study temporal relations and event compositions, respectively, while Omni-Embed-Audio (Yoo et al., 2026) investigates robustness to queries expressing different user intents.

For fine-grained alignment, one line enriches clip–caption supervision: FineLAP (Li et al., 2026b) combines clip- and frame-level supervision, while FLAM (Wu et al., 2025) obtains frame-wise supervision using simulated audio and generated captions. Another line refines matching represen-

tations: GPA (Xie et al., 2024) aligns global and prototype representations, MGA-CLAP (Li et al., 2024) uses a shared codebook and locality-aware representations for frame–word correspondence, and THA+DCR (Xin et al., 2024) combines hierarchical alignment with disentangled cross-modal representations. When retained at inference, local matching introduces computation beyond a single global similarity score. SAYAKA uses clip–caption pairs for concept-level supervision without additional temporal annotations and retains only global embeddings for retrieval.

Alongside alignment design, contrastive learning also studies the construction of training candidate sets. MoCo (He et al., 2020) uses a momentum encoder and a queue to extend comparisons beyond the current batch, while hard-negative sampling (Robinson et al., 2021) focuses on more confusable candidates. For audio–text retrieval, Xie et al. (2023) report benefits from cross-modal semi-hard negatives over hard negatives. SAYAKA uses these mechanisms to support global and concept-level training, without changing its inference procedure.

3 METHOD

3.1 OVERVIEW

SAYAKA decomposes captions into semantic concepts and uses each concept to query a shared audio-token sequence, extracting concept-specific acoustic representations. Graded alignment and concept separation constraints then model weak-positive relationships and distinctions among concepts in the same caption. Adaptive recomposition transfers concept-level training signals to the global representation used at inference, while multi-granularity momentum queues extend contrast beyond the batch to strengthen discrimination. Concept extraction relies only on parsing the paired captions, without temporal annotations or external data synthesis. The concept branch is used only during training; at inference, SAYAKA reduces to a standard dual encoder and retains the retrieval efficiency of global audio–text retrieval (ATR).

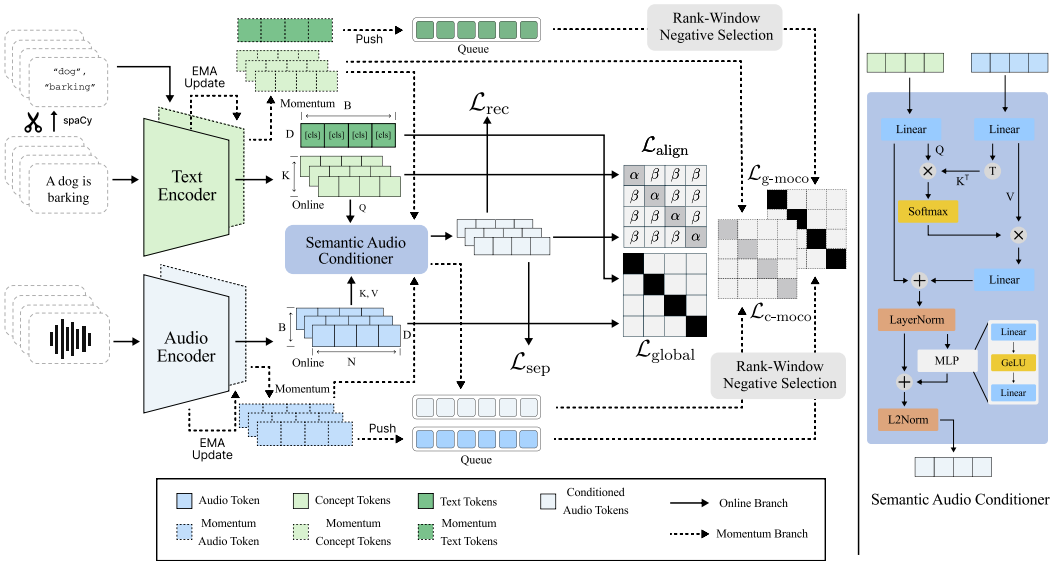


Figure 1: SAYAKA overview. Online and momentum branches learn global and concept-level representations with Rank-Window Negative Selection (left); semantic attention yields Conditioned Audio Tokens (right). Momentum encoders and queues are training-only; inference retrieves by cosine similarity of global audio and text embeddings.

3.2 AUDIO–CONCEPT REPRESENTATION LEARNING

Concept Extraction and Encoding. Audio captions describe multiple entities, actions, and relations, but clip-level pairs provide no temporal boundaries, and sounds may overlap. Without explicit temporal boundaries, frame-level alignment is difficult; we therefore approach the problem from

the text side by decomposing captions into concepts. We use spaCy to extract noun phrases, action phrases, and spatial phrases from caption t_i, forming $\mathcal{C}_i = \{c_{ik}\}_{k=1}^{K_i}$ after deduplication. Each of the K_i valid concepts queries the full audio-token sequence, extracting concept-relevant information without predefined temporal segmentation.

For a batch of B audio–caption pairs, a CED audio encoder (Dinkel et al., 2024) encodes audio, while a BERT text encoder (Devlin et al., 2019) encodes both captions and concepts. We denote the local audio tokens by H_i and concept features by E_{ik}. The projected, L_2-normalized global audio and caption representations A_i and T_i lie in a shared d-dimensional space and are aligned with the bidirectional contrastive loss $\mathcal{L}_{\text{global}}$ (Oord et al., 2018).

Semantically Conditioned Audio. Text conditioning directs audio representations toward query-relevant content (Li et al., 2026a). Here, each parsed concept provides the conditioning signal for extracting its corresponding acoustic evidence from the shared audio sequence. Semantic attention takes concept features as queries and audio tokens as keys and values, enabling this directed extraction through cross-attention (Vaswani et al., 2017). We project the concept features and local audio sequence into a shared dimension, obtaining queries $q_{ik} = P_c E_{ik}$ and an audio sequence $X_i = P_a H_i$, where P_c and P_a are learnable projections:

$$\begin{aligned} R_{ik} &= \mathrm{LN}(\mathrm{MHA}(q_{ik}, X_i, X_i) + q_{ik}), \\ U_{ik} &= \mathrm{L2Norm}(R_{ik} + \mathrm{MLP}(R_{ik})). \end{aligned} \tag{1}$$

The query residual preserves the semantic condition while aggregating acoustic content, yielding conditioned audio tokens U_{ik}.

Soft-Positive Audio–Concept Alignment. Concepts within the same caption jointly describe an audio clip and are semantically related. Treating all non-matching concepts as negatives would ignore these relationships. We therefore adopt Soft-Positive Audio–Concept Alignment: each conditioned audio token receives weight α for its corresponding concept, weight β for other concepts in the same caption, and zero weight for concepts from other captions, with $\alpha = 1.0$ and $\beta = 0.5$ in the main configuration. This emphasizes direct correspondence while preserving semantic relationships within the clip (Khosla et al., 2020). Let $C_{ik} = \mathrm{L2Norm}(q_{ik})$, and flatten the valid pairs in the batch into $\{(U_u, C_u)\}_{u=1}^{V}$, where $V = \sum_i K_i$. The audio–concept similarity is $S_{uv} = U_u^\top C_v$. We construct a weight matrix W using the rule above and normalize each row to obtain the target distribution $D_{uv} = W_{uv} / \sum_z W_{uz}$. The bidirectional graded alignment objective is

$$\mathcal{L}_{\text{align}} = \frac{1}{2}\left[\mathrm{CE}_{\text{soft}}(s_c S, D) + \mathrm{CE}_{\text{soft}}(s_c S^\top, D)\right]. \tag{2}$$

The learnable positive scale s_c adjusts the logits, and $\mathrm{CE}_{\text{soft}}$ averages soft-target cross-entropy over rows. The weighting rule yields the same target distribution D in both directions.

Audio-Token Separation. Graded alignment preserves semantic relationships within a caption, but audio representations conditioned on different concepts may still aggregate similar content. To reduce this redundancy, we impose a separation constraint on representations of distinct concepts within each clip:

$$\mathcal{L}_{\text{sep}} = \frac{1}{|\mathcal{P}|} \sum_{(i,k,\ell)\in\mathcal{P}} \exp\left[\gamma\left(U_{ik}^\top U_{i\ell} - \delta\right)\right]. \tag{3}$$

The sum runs over ordered pairs of distinct valid concepts within each clip, collected in $\mathcal{P}$. The penalty grows with similarity at a rate controlled by $\gamma > 0$, with δ as a shift constant.

Global Consistency of Audio Representations. Concept-conditioned representations exist only during training and do not participate in inference scoring. To transfer their information to the global audio embedding, we introduce a recomposition objective related to reconstruction-based supervision in audio–text retrieval (Luo et al., 2023). This objective fuses the conditioned representations into a single vector and aligns it with the global audio embedding used at inference. A learnable scalar scoring function r assigns weights $w_{ik} = \mathrm{softmax}_k(r(U_{ik}))$ over valid concepts. We aggregate $\bar{A}_i = \sum_k w_{ik} U_{ik}$ and recompose $\hat{A}_i = \mathrm{L2Norm}(\bar{A}_i + \mathrm{MLP}_r(\bar{A}_i))$ through a residual mapping. The bidirectional contrastive objective is

$$\begin{aligned} G_{ij}^{\text{rec}} &= s_g\, \hat{A}_i^\top A_j, \\ \mathcal{L}_{\text{rec}} &= \frac{1}{2}\left[\mathrm{CE}(G^{\text{rec}}, y) + \mathrm{CE}((G^{\text{rec}})^\top, y)\right]. \end{aligned} \tag{4}$$

Labels $y_i = i$ match representations of the same clip, and the objective shares the scale s_g with global contrast. This ensures that the global encoder aligns with fine-grained concept information during training.

3.3 Momentum Queue Learning

Discriminative Role of Hard Negatives. A momentum encoder (He et al., 2020) uses an exponential moving average to generate stable keys stored in first-in, first-out queues, extending comparisons beyond the batch. Expanding the queue increases the candidate pool, but many easy negatives can dilute the relative contribution of hard negatives. Rank-Window Negative Selection sorts valid queued keys by cosine similarity to each query in descending order and retains the top K_{neg} candidates. This focuses cross-batch supervision on harder negatives (Robinson et al., 2021) while limiting the influence of false negatives (Chuang et al., 2020). Globally, online audio queries are paired with momentum text keys and contrasted against selected text-queue negatives; the reverse direction uses the audio queue. Averaging the two InfoNCE losses gives $\mathcal{L}_{\text{g-moco}}$.

Concept Repetition Analysis. Concept-level contrast must account for repeated phrases across captions. On the AudioCaps training set, spaCy yields 137,318 concept occurrences but only 23,899 distinct texts. After excluding candidates from the same audio, nearly 90% of queries have an exact text match at rank 1. In audio-to-concept contrast, repeated texts may act as both positives and queued negatives, and top-K_{neg} selection may further concentrate these conflicts.

Concept-to-Audio Contrast. We therefore retain only concept-to-audio contrast, using online concepts as queries and full momentum conditioned audio representations as keys. Their clip-specific acoustic content supports instance discrimination, avoiding the reverse objective's conflicting supervision over repeated text keys.

Let $\mathcal{I}$ index valid concepts, U_{ik}^{+} denote the momentum conditioned audio key paired with C_{ik}, and $\mathcal{N}_{ik}$ contain the selected conditioned audio negatives. The concept-level objective is

$$\mathcal{L}_{\text{c-moco}} = -\frac{1}{|\mathcal{I}|} \sum_{(i,k)\in\mathcal{I}} \log \frac{\exp(s_c C_{ik}^{\top} U_{ik}^{+})}{\exp(s_c C_{ik}^{\top} U_{ik}^{+}) + \sum_{U^{-}\in\mathcal{N}_{ik}} \exp(s_c C_{ik}^{\top} U^{-})}. \tag{5}$$

Here, s_c is shared with in-batch graded alignment, and momentum keys receive no gradients. Only concept-level momentum contrast is unidirectional; in-batch graded alignment and global contrast remain bidirectional.

3.4 Training and Inference

The complete training objective combines global retrieval supervision, fine-grained structural constraints, and cross-batch contrast:

$$\begin{aligned}\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{global}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} \\ + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{g-moco}}\mathcal{L}_{\text{g-moco}} + \lambda_{\text{c-moco}}\mathcal{L}_{\text{c-moco}}.\end{aligned} \tag{6}$$

Each λ weights the corresponding auxiliary loss. At inference, only the global representations $A(x)$ and $T(t)$ are retained and ranked by cosine similarity. The concept-conditioned path, momentum encoder, and queues are no longer needed, so SAYAKA preserves the independent encoding and retrieval efficiency of the CLAP bi-encoder.

4 Experiments

We evaluate SAYAKA through retrieval comparisons, progressive component ablations, and analyses of momentum queues and concept-conditioned representations. The experiments examine whether concept-level training improves global retrieval, how the queue design affects performance, and how different concepts condition representations of the same audio clip.

4.1 Experimental Setup

Datasets. We evaluate on AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020), two standard benchmarks differing in recording source, duration, and caption style. AudioCaps pairs ap-

proximately 10-second YouTube clips from AudioSet (Gemmeke et al., 2017), covering diverse real-world sounds, with crowd-written descriptions. Clotho contains 15–30-second Freesound recordings, each with five human-written captions of 8–20 words.

Evaluation Metrics. We report Recall@1, Recall@5, and mean Average Precision at 10 (mAP@10) for text-to-audio (T2A) and audio-to-text (A2T) retrieval. Recall@k measures the fraction of queries with a relevant result among the top k candidates. For mAP@10, precision at relevant positions in the top ten is summed, divided by the total number of relevant candidates, and averaged over queries. We report percentages (higher is better) as three-run means and sample standard deviations.

Implementation Details. We fine-tune pretrained CED-Base (Dinkel et al., 2024) and BERT-Base-Uncased (Devlin et al., 2019) encoders with 1024-dimensional embeddings, eight-head cross-attention, and at most 30 concepts per clip. Loss weights are $\lambda_{\text{align}} = \lambda_{\text{g-moco}} = \lambda_{\text{c-moco}} = 1.0$ and $\lambda_{\text{sep}} = \lambda_{\text{rec}} = 0.1$. Both datasets use Adam (Kingma & Ba, 2015) for 20 epochs with batch size 24, zero weight decay, and an initial learning rate of 5×10^{-5} reduced tenfold every four epochs. Training uses bfloat16 on NVIDIA GeForce RTX 4090 GPUs with seeds 41, 42, and 6657; checkpoints are selected by Recall@1. EMA decay is $m = 0.999$; global MoCo uses temperature 0.07, while concept-level MoCo shares the learnable alignment scale initialized to $1/0.07$. Queue settings are analyzed in Appendix C; alignment, separation, and recomposition hyperparameters in Appendix A.

4.2 Main Results

Comparison Methods. We compare CLAP (Elizalde et al., 2023), ML-ACT (Mei et al., 2022), THA+DCR (Xin et al., 2024), GPA (Xie et al., 2024), MGA-CLAP (Li et al., 2024), FLAM (Wu et al., 2025), m-LTM (Luong et al., 2024), DTU (Oncescu et al., 2024), SupCLAP (Luo et al., 2026), and Cacophony (Zhu et al., 2024). Within each dataset, all compared methods are retrained using the same CED-Base and BERT-Base-Uncased backbone as SAYAKA. This controls backbone capacity and reduces confounding from encoder differences. CLAP is the global dual-encoder baseline.

Table 1: Recall on AudioCaps and Clotho. All methods use CED-Base and BERT-Base-Uncased. Values are percentages (mean ± sample standard deviation); bold marks the highest reported mean.

Method	AudioCaps				Clotho			
	A2T		T2A		A2T		T2A	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLAP	48.03 ± 0.70	80.42 ± 0.64	38.23 ± 0.25	75.33 ± 0.20	20.08 ± 0.51	45.37 ± 2.09	15.75 ± 0.20	42.30 ± 0.71
ML-ACT	47.51 ± 0.94	80.36 ± 0.28	39.29 ± 0.10	75.13 ± 0.34	19.49 ± 0.87	44.55 ± 0.44	15.93 ± 0.84	42.02 ± 0.61
THA+DCR	51.93 ± 3.98	83.42 ± 1.87	41.75 ± 1.89	77.32 ± 0.98	19.20 ± 0.63	43.75 ± 0.58	15.96 ± 0.66	41.26 ± 0.08
GPA	49.49 ± 1.11	81.85 ± 1.33	39.64 ± 0.26	76.38 ± 0.29	20.80 ± 1.45	45.51 ± 0.45	16.33 ± 0.47	42.65 ± 0.27
MGA-CLAP	50.37 ± 0.28	81.71 ± 0.73	42.15 ± 0.40	**78.02 ± 0.31**	21.96 ± 0.95	46.25 ± 1.62	17.47 ± 0.44	43.55 ± 0.61
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
m-LTM	53.61 ± 0.21	**84.26 ± 0.91**	40.04 ± 0.92	75.89 ± 0.63	22.15 ± 1.40	47.85 ± 0.06	17.24 ± 0.85	42.86 ± 0.92
DTU	46.50 ± 2.08	79.45 ± 0.77	39.11 ± 0.57	75.06 ± 0.31	19.78 ± 1.22	43.62 ± 1.37	15.60 ± 0.60	41.53 ± 0.64
SupCLAP	49.04 ± 0.12	80.46 ± 0.38	39.75 ± 0.43	75.85 ± 0.31	21.09 ± 0.81	44.68 ± 1.15	16.31 ± 0.41	41.93 ± 0.15
Cacophony	49.63 ± 2.23	81.23 ± 0.74	39.00 ± 0.61	75.74 ± 0.15	19.36 ± 0.48	43.85 ± 1.09	16.61 ± 0.42	42.00 ± 0.68
SAYAKA (Ours)	**55.49 ± 0.10**	84.15 ± 0.97	**43.01 ± 0.24**	77.83 ± 0.55	**22.32 ± 0.29**	**48.45 ± 1.39**	**17.70 ± 0.64**	**44.85 ± 0.35**

Results. SAYAKA achieves the highest Recall@1 means in both directions on both datasets and leads nine of twelve metrics in Tables 1 and 2. AudioCaps T2A/A2T Recall@1 reaches 43.01/55.49, exceeding CLAP by 4.78/7.45 points and the strongest respective competitors by 0.86/1.88 points. T2A/A2T mAP@10 also leads at 57.51/41.06 versus CLAP's 53.76/35.98; both Recall@5 means are within 0.20 points of the best.

On Clotho, SAYAKA leads all four Recall metrics, with A2T/T2A Recall@1 of 22.32/17.70. T2A mAP@10 leads at 29.25; A2T mAP@10 is 15.66, only 0.03 points below the best. These results show consistent gains over CLAP using only global audio–text cosine similarity, supporting the value of concept-level training for global bi-encoder retrieval.

4.3 Retrieval Efficiency Analysis

SAYAKA uses only global audio and text embeddings at retrieval time; Semantically Conditioned Audio and momentum queues are training-only. Table 3 compares matching, ranking, and their

Table 2: mAP@10 on AudioCaps and Clotho. Values are percentages (mean ± sample standard deviation); bold marks the highest reported mean.

Method	AudioCaps		Clotho	
	A2T	T2A	A2T	T2A
CLAP	35.98 ± 0.23	53.76 ± 0.20	14.11 ± 0.46	27.08 ± 0.43
ML-ACT	35.98 ± 0.14	54.38 ± 0.15	14.09 ± 0.37	27.14 ± 0.61
THA+DCR	39.22 ± 2.56	56.80 ± 1.62	14.06 ± 0.21	26.79 ± 0.39
GPA	37.28 ± 0.26	54.99 ± 0.09	15.00 ± 0.16	27.73 ± 0.22
MGA-CLAP	38.61 ± 0.11	57.12 ± 0.15	$\mathbf{15.69 \pm 0.50}$	28.57 ± 0.36
FLAM	37.13 ± 0.49	53.75 ± 0.21	14.57 ± 0.10	27.29 ± 0.31
m-LTM	39.40 ± 0.99	55.08 ± 0.80	15.45 ± 0.55	28.24 ± 0.49
DTU	35.12 ± 0.06	54.22 ± 0.34	14.04 ± 0.30	26.84 ± 0.57
SupCLAP	36.57 ± 0.11	54.87 ± 0.36	14.73 ± 0.31	27.41 ± 0.12
Cacophony	36.96 ± 0.53	54.44 ± 0.37	14.37 ± 0.37	27.56 ± 0.39
SAYAKA (Ours)	$\mathbf{41.06 \pm 0.35}$	$\mathbf{57.51 \pm 0.14}$	15.66 ± 0.48	$\mathbf{29.25 \pm 0.52}$

totals for CLAP and fine-grained methods. CLAP, SAYAKA, and FLAM share the lowest matching costs and totals among the methods shown, while m-LTM, GPA, and THA+DCR incur greater matching costs. Appendix D reports stage totals and CLAP-relative ratios for more methods.

Table 3: Matching, ranking, and total times (ms; lower is better); bold marks minimum totals. CLAP, SAYAKA, and FLAM use updated three-decimal measurements. The m-LTM AudioCaps total sums displayed stages; GPA's reported total differs by 0.01 ms from the rounded stage sum.

	AudioCaps			Clotho		
Method	Match	Rank	Total	Match	Rank	Total
CLAP	0.291	0.973	**1.264**	0.315	1.353	**1.668**
SAYAKA	0.291	0.973	**1.264**	0.315	1.353	**1.668**
FLAM	0.291	0.973	**1.264**	0.315	1.353	**1.668**
m-LTM	3.96	0.98	4.94	4.21	1.29	5.50
GPA	286.79	1.00	287.80	366.51	1.37	367.88
THA+DCR	177,085.95	0.98	177,086.93	222,802.21	1.35	222,803.56

m-LTM's matching stage takes 3.96 ms on AudioCaps versus 0.291 ms for CLAP and SAYAKA. GPA and THA+DCR have still larger matching costs, whereas ranking times vary much less. SAYAKA retains CLAP's matching and ranking costs while using fine-grained concept supervision during training.

4.4 Ablation Studies

We evaluate SAYAKA through individual and grouped objective removals on AudioCaps (Table 4).

Concept-Level Objectives. Relative to CLAP with Global Audio-Text MoCo, full SAYAKA improves A2T and T2A Recall@1 means by 3.49 and 1.32 points. This comparison holds global momentum supervision in place, indicating that the concept-level objectives provide additional benefits beyond the global contrastive objective. Removing separation or recomposition lowers A2T Recall@1 by 1.36 or 1.75 points, respectively, while increasing T2A Recall@1 by 0.29 or 0.41 points. Separation discourages redundant Conditioned Audio Tokens, whereas recomposition constrains their aggregate to remain consistent with the global audio representation. The larger A2T changes are consistent with these constraints helping an audio query distinguish among captions describing overlapping aspects of a scene. However, the smaller T2A changes show that neither constraint improves every retrieval direction in isolation. The full objective therefore favors a balance between concept differentiation and global consistency, with the clearest benefit observed in A2T retrieval.

Table 4: AudioCaps ablations (%, mean ± sample standard deviation). The last row retains only CLAP and Global Audio-Text MoCo; other w/o rows remove the named loss from full SAYAKA. Bold denotes the highest mean in each column.

Method	A2T	T2A
	R@1	R@1
SAYAKA (full)	**55.49 ± 0.10**	43.01 ± 0.24
w/o Audio-Token Separation Loss	54.13 ± 1.06	43.30 ± 0.44
w/o Audio Recomposition Consistency Loss	53.74 ± 0.63	**43.42 ± 0.55**
w/o Concept-Level Audio-Text MoCo	54.27 ± 0.54	42.52 ± 0.36
w/o Concept-Level Objectives	52.00 ± 1.36	41.69 ± 0.52

Momentum Objectives. Removing Concept-Level Audio-Text MoCo lowers A2T and T2A Recall@1 means by 1.22 and 0.49 points. Compared with the last row, which retains only CLAP and Global Audio-Text MoCo, full SAYAKA gains 3.49 and 1.32 points, respectively. Thus, the concept-level objectives improve retrieval beyond global momentum contrast, with the concept queue contributing further gains.

Contrastive Direction. Repeated concept phrases can give different audio clips identical text keys, making them ambiguous negatives for audio-to-concept (A2C) contrast. Concept-to-audio (C2A) contrast instead uses Conditioned Audio Tokens that incorporate each clip's acoustic content. We compare the two directions in Table 5 and analyze concept repetition in Section 4.5.

Table 5: AudioCaps ablation of concept-level momentum direction (%, mean ± sample standard deviation). Capacity 2400 and Top-24 are fixed; global momentum remains bidirectional.

Concept-level direction	A2T	T2A
	R@1	R@1
Bidirectional	54.82 ± 0.70	43.01 ± 0.72
C2A-only	55.49 ± 0.10	43.01 ± 0.24
A2C-only	53.26 ± 0.34	42.51 ± 0.67

Holding capacity at 2400 and retaining ranks 1–24, C2A-only achieves the highest A2T Recall@1, exceeding bidirectional contrast by 0.66 points while matching its T2A mean (Table 5). It also outperforms A2C-only in both directions. Global momentum contrast remains bidirectional in all three settings. The results favor C2A; Section 4.5 examines the repeated text keys.

4.5 Concept-Level Insights

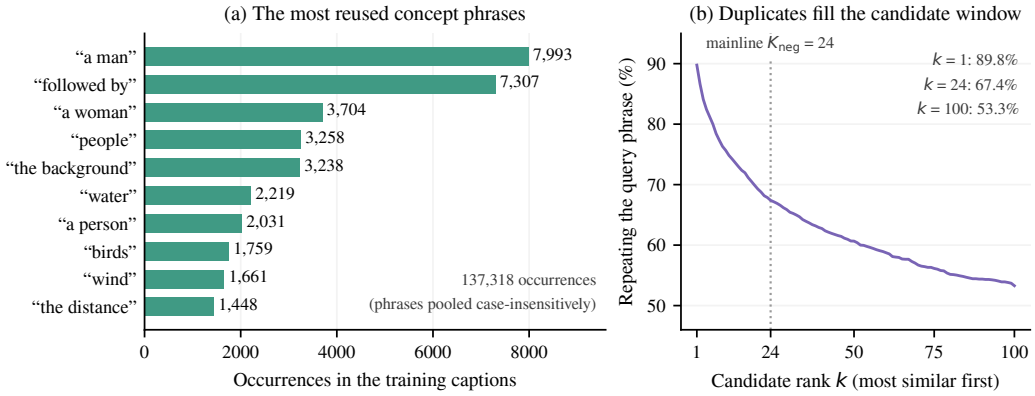


Figure 2: Concept repetition in AudioCaps training captions. (a) Most reused phrases. (b) Fraction of rank-k neighbors repeating the query text; dotted line: the $K_{\text{neg}} = 24$ negative-selection window.

Parsed Concepts. AudioCaps captions contain an average of 2.79 concepts in training and 3.54 in testing; 79.42% and 84.90%, respectively, contain at least two concepts. Noun phrases and actions dominate, providing multiple semantic queries about entities and actions for the shared audio sequence. Appendix B gives the full count and category distributions.

Concept Repetition. To understand the direction ablation in Table 5, we examine repetition among text-concept candidates. AudioCaps training captions contain 137,318 concept occurrences but only 23,899 distinct texts (Figure 2(a)). Excluding candidates from the same audio, exact text matches account for 89.76%, 67.40%, and 53.28% of candidates at ranks 1, 24, and 100, respectively (Figure 2(b)). Repeated phrases thus occupy many highly ranked candidate positions across recordings, making text keys ambiguous negatives for A2C contrast. This observation supports the C2A choice, which retains clip-specific acoustic information in its candidate keys.

Concept-Conditioned Attention. Figure 3 examines the caption "A person is typing in a computer keyboard," whose parsed concepts are "A person," "a computer keyboard," and "typing in." We visualize the best checkpoint of the main configuration. The three concepts query the same audio sequence to capture different semantic aspects of the typing event.

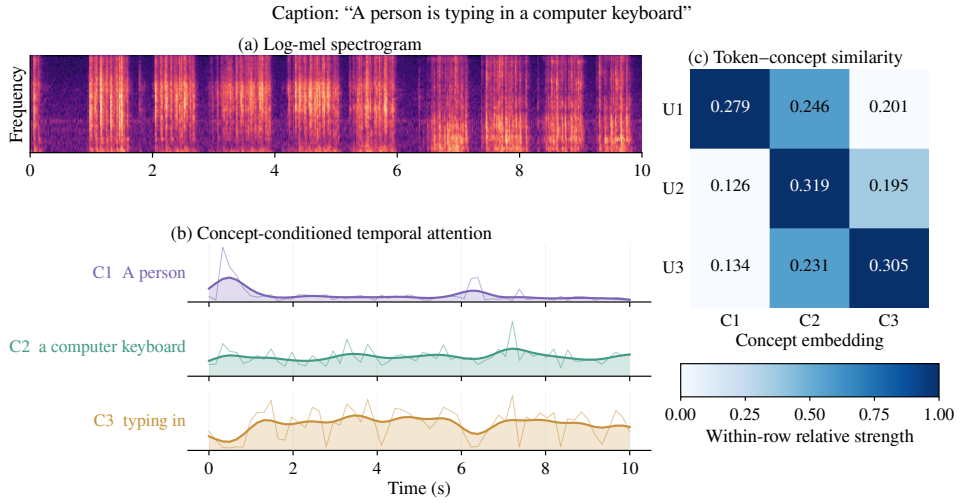


Figure 3: Concept-conditioned keyboard-typing example. (a) Log-mel spectrogram. (b) Attention over 0–10 s, normalized by each row's peak; thin traces are raw, filled curves Gaussian-smoothed ($\sigma = 2$ bins). (c) Token–concept similarity: row-normalized colors and raw cosine values.

Against the spectrogram, keyboard- and action-conditioned attention extends across much of the clip, whereas person-conditioned attention concentrates near the beginning. Each Conditioned Audio Token also has its highest cosine similarity with the corresponding concept. Together, these patterns show distinct but overlapping semantic responses to the same acoustic event.

5 Conclusion

In this paper, we presented SAYAKA, a fine-grained audio-text retrieval framework that introduces concept-level alignment without changing the bi-encoder inference paradigm. Unlike approaches that rely on additional temporal supervision or retain local matching at inference time, SAYAKA exploits semantic concepts directly parsed from paired captions and uses them only during training to decompose and recompose concept-relevant acoustic evidence from shared audio representations. In this way, fine-grained supervision is ultimately transferred to the global retrieval embedding used at inference. Experimental results show that this design achieves a favorable balance between retrieval accuracy and efficiency, and further suggests that the internal concept structure of open-vocabulary captions can itself serve as an effective supervision source for fine-grained alignment. We hope SAYAKA offers a lightweight and scalable direction for fine-grained audio-text retrieval, and provides useful insight for future work on richer language structures and more complex auditory scenes in cross-modal alignment.

AI USE STATEMENT

During the preparation of this manuscript, we used generative AI tools to assist with language polishing and LaTeX writing. Specifically, these tools helped improve the clarity and flow of selected sentences, suggested wording changes, and assisted with LaTeX typesetting and formatting. Their role in these tasks was limited to providing suggestions for the authors to review; they did not serve as a source of research data. We did not use generative AI to generate or synthesize any datasets used in this study, nor to create synthetic audio, text descriptions, or annotations.

All suggested textual revisions and LaTeX edits were checked by the authors against the experimental records before being incorporated into the manuscript. The authors also verified that the relevant descriptions accurately reflect the methods actually used, the experimental settings, and the research results. The authors are responsible for the paper's scientific claims, interpretation of results, and final presentation, and take full responsibility for its content.

REFERENCES

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020.

Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. Audio Retrieval with WavText5K and CLAP Training. In *Interspeech 2023*, pp. 2948–2952, 2023. doi: 10.21437/Interspeech.2023-1136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT (Volume 1)*, pp. 4171–4186, 2019.

Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang. CED: Consistent ensemble distillation for audio tagging. In *Proceedings of ICASSP*, pp. 291–295. IEEE, 2024.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proceedings of 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 *Proceedings of ICASSP*, pp. 1–5. IEEE, 2023.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of ICASSP*, pp. 776–780. IEEE, 2017.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. CompA: Addressing the gap in compositional reasoning in audio-language models. In *Proceedings of ICLR*, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*, pp. 9726–9735. IEEE, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of NAACL-HLT (Volume 1)*, pp. 119–132, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of ICLR*, 2015.

Mohan Li, Rama Doddipatla, and Philip C. Woodland. Text-Prompted CLAP: Learning text-conditioned audio representations via contrastive learning. arXiv preprint arXiv:2607.25085, 2026a. URL https://arxiv.org/abs/2607.25085.

Xiquan Li, Xuenan Xu, Ziyang Ma, Wenxi Chen, Haolin He, Qiuqiang Kong, and Xie Chen. FineLAP: Taming heterogeneous supervision for fine-grained language-audio pretraining. In *Proceedings of ACL*, pp. 10393–10408, 2026b.

Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. Advancing multi-grained alignment for contrastive language-audio pre-training. In *Proceedings of ACM Multimedia*, pp. 7356–7365, 2024.

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. In *Proceedings of ICLR*, 2026.

Kaiyi Luo, Xulong Zhang, Jianzong Wang, Huaxiong Li, Ning Cheng, and Jing Xiao. Contrastive latent space reconstruction learning for audio-text retrieval. In *Proceedings of ICTAI*, pp. 913–917, 2023. doi: 10.1109/ICTAI59109.2023.00137.

Tien Manh Luong, Khai Nguyen, Nhat Ho, Reza Haffari, Dinh Phung, and Lizhen Qu. Revisiting deep audio-text retrieval through the lens of transportation. In *Proceedings of ICLR*, 2024.

Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang. On Metric Learning for Audio-Text Cross-Modal Retrieval. In *Interspeech 2022*, pp. 4142–4146, 2022. doi: 10.21437/Interspeech.2022-11115.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 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, 2024.

Andreea-Maria Oncescu, João F. Henriques, and A. Sophia Koepke. Dissecting temporal understanding in text-to-audio retrieval. In *Proceedings of ACM Multimedia*, pp. 9535–9543, 2024.

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

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *Proceedings of ICLR*, 2021.

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

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 *Proceedings of 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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67719–67740. PMLR, 13–19 Jul 2025.

Huang Xie, Okko Räsänen, and Tuomas Virtanen. On negative sampling for contrastive audio-text retrieval. In *Proceedings of ICASSP*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095319.

Yuxin Xie, Zhihong Zhu, Xianwei Zhuang, Liming Liang, Zhichang Wang, and Yuexian Zou. GPA: Global and prototype alignment for audio-text retrieval. In *Interspeech 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. In *Interspeech 2024*, pp. 1140–1144, 2024. doi: 10.21437/Interspeech.2024-593.

HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo, and Du-Seong Chang. Omni-Embed-Audio: Leveraging multimodal LLMs for robust audio-text retrieval. In *Proceedings of ACL (Volume 1: Long Papers)*, pp. 22664–22680, July 2026. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1038.

Ge Zhu, Jordan Darefsky, and Zhiyao Duan. Cacophony: An improved contrastive audio-text model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4867–4879, 2024.

A Hyperparameter Ablations

We vary one hyperparameter at a time on AudioCaps while holding the remaining settings fixed. Table 6 reports the mean and sample standard deviation over three runs. Bold metric values denote the highest mean within each hyperparameter block. Bold hyperparameter values denote the settings retained in the main configuration. Here, β is the graded alignment weight; λ_{sep}, δ, and γ are the weight, margin, and scale of Audio-Token Separation; and λ_{rec} is the Audio Recomposition Consistency weight.

Table 6: Hyperparameter ablations on AudioCaps. Values are percentages (mean ± sample standard deviation). Bold metric values indicate the highest mean in each block; bold hyperparameter values are retained in the main configuration.

Hyperparameter	**A2T**		**T2A**	
	R@1	R@5	R@1	R@5
$\beta = 0.0$	52.91 ± 0.84	$\mathbf{84.71 \pm 0.32}$	42.70 ± 0.15	76.88 ± 0.34
$\beta = 0.25$	54.41 ± 0.74	84.57 ± 0.47	43.58 ± 0.14	77.45 ± 0.35
$\beta = \mathbf{0.5}$	$\mathbf{55.49 \pm 0.10}$	84.15 ± 0.97	43.01 ± 0.24	$\mathbf{77.83 \pm 0.55}$
$\beta = 0.75$	53.88 ± 0.58	84.47 ± 1.00	43.32 ± 0.77	77.50 ± 0.21
$\beta = 1.0$	54.44 ± 0.73	84.05 ± 0.34	$\mathbf{43.63 \pm 0.55}$	77.23 ± 0.45
$\lambda_{\text{sep}} = 0.0$	54.13 ± 1.06	84.33 ± 0.65	43.30 ± 0.44	77.26 ± 0.35
$\lambda_{\text{sep}} = 0.05$	53.61 ± 0.10	84.67 ± 0.37	$\mathbf{43.33 \pm 0.33}$	77.11 ± 0.55
$\lambda_{\text{sep}} = \mathbf{0.1}$	$\mathbf{55.49 \pm 0.10}$	84.15 ± 0.97	43.01 ± 0.24	$\mathbf{77.83 \pm 0.55}$
$\lambda_{\text{sep}} = 0.2$	54.58 ± 0.49	$\mathbf{84.74 \pm 0.72}$	43.21 ± 0.10	77.29 ± 0.14
$\delta = 0.4$	53.54 ± 0.52	$\mathbf{84.36 \pm 1.35}$	$\mathbf{43.29 \pm 0.34}$	76.90 ± 0.10
$\delta = \mathbf{0.6}$	$\mathbf{55.49 \pm 0.10}$	84.15 ± 0.97	43.01 ± 0.24	$\mathbf{77.83 \pm 0.55}$
$\delta = 0.8$	54.89 ± 0.42	84.33 ± 0.18	43.23 ± 0.34	77.23 ± 0.74
$\gamma = 0.25$	53.61 ± 0.28	84.29 ± 0.42	43.13 ± 0.35	77.28 ± 0.06
$\gamma = \mathbf{0.5}$	$\mathbf{55.49 \pm 0.10}$	84.15 ± 0.97	43.01 ± 0.24	$\mathbf{77.83 \pm 0.55}$
$\gamma = 1.0$	54.62 ± 0.71	$\mathbf{84.36 \pm 0.73}$	$\mathbf{43.67 \pm 0.53}$	77.56 ± 0.32
$\lambda_{\text{rec}} = 0.0$	53.74 ± 0.63	84.01 ± 0.38	43.42 ± 0.55	77.29 ± 0.07
$\lambda_{\text{rec}} = 0.05$	54.37 ± 1.15	83.77 ± 1.07	$\mathbf{43.44 \pm 0.21}$	77.04 ± 0.47
$\lambda_{\text{rec}} = \mathbf{0.1}$	$\mathbf{55.49 \pm 0.10}$	84.15 ± 0.97	43.01 ± 0.24	$\mathbf{77.83 \pm 0.55}$
$\lambda_{\text{rec}} = 0.2$	54.06 ± 0.64	$\mathbf{84.05 \pm 0.52}$	43.41 ± 0.41	77.43 ± 0.27

Concepts in the same audio clip often share acoustic context without referring to identical content. Soft-Positive Audio–Concept Alignment therefore assigns weight α to the corresponding concept and a smaller weight β to other concepts in the same caption. This avoids treating co-occurring concepts as negatives while prioritizing the direct correspondence. After normalization, the weights also determine how supervision is distributed across concepts: for a caption with K valid concepts, the target mass on the corresponding concept is $\alpha/[\alpha + (K-1)\beta]$. Thus, β controls both the relationship among concepts and the emphasis on each conditioned audio token's own concept as the number of concepts increases.

Soft-positive alignment structures the relationship between audio representations and text concepts, but does not explicitly constrain pairwise similarity among conditioned audio tokens from the same clip. Audio-Token Separation supplies this constraint, allowing the tokens to retain shared acoustic context while encoding different concept-specific information. Its exponential loss penalizes more

similar concept pairs more strongly: γ controls sensitivity to differences in pairwise similarity, while λ_{sep} sets the contribution of this objective to the joint loss. For the loss used here, changing δ at fixed γ rescales the separation term rather than creating a hard threshold below which the penalty vanishes. Together, these parameters regulate the separation pressure alongside the semantic relationships preserved by soft-positive alignment.

Distinct concept-level representations must also inform the retrieval representation, since conditioned audio tokens are not used at inference. Adaptive Audio Recomposition learns weights over these tokens, aggregates them into a recomposed audio representation, and aligns that representation with the global audio embedding through bidirectional contrastive consistency. This objective pairs the recomposed and global representations of the same audio clip while distinguishing other clips in the batch, bringing aggregated concept information into global instance discrimination. The weight λ_{rec} controls this connection between fine-grained training signals and the global audio representation used for cosine-similarity retrieval with global text representations.

B Concept Parsing Statistics

The training and test sets contain 49,274 and 4,669 captions, respectively. Figure 4(a) shows that parsed concept counts range from one to multiple concepts, with the test distribution shifted toward larger sets. Each caption thus yields a variable-sized concept set: every concept queries the same audio-token sequence rather than a predefined audio segment. The $K \leq 30$ limit exceeds the largest parsed concept sets observed in these AudioCaps splits (15 in training and 14 in testing), retaining all extracted concepts without truncation. This variation also changes the training relationships within a sample. A caption with K valid concepts provides K concept queries and their corresponding audio–concept matches, alongside soft-positive relationships with other concepts from the same caption. Its conditioned audio representations also form $K(K-1)$ ordered pairs of distinct concepts for Audio-Token Separation. Captions with more concepts therefore supply more queries and more relationships through which to learn conceptual relatedness and representation-level distinction.

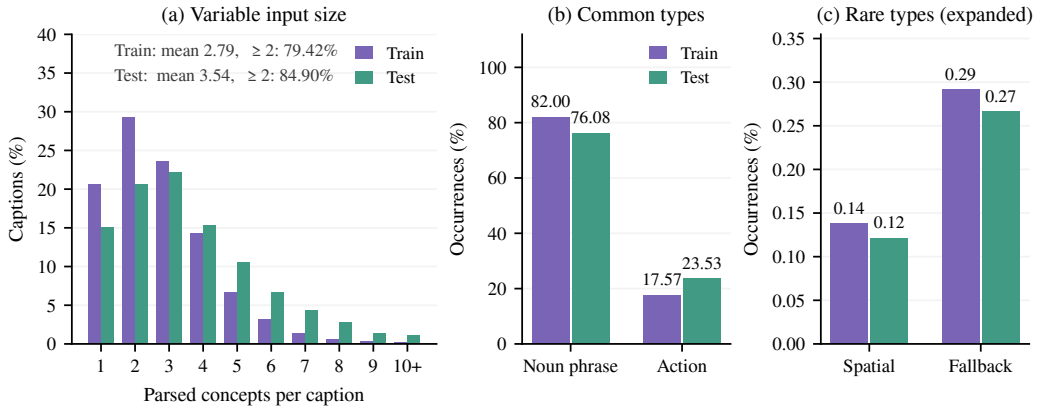


Figure 4: Distribution of parsed concept counts and concept types in AudioCaps. (a) Captions per parsed concept count; the panel reports the mean count and the share of captions with at least two concepts. (b)–(c) Share of concept occurrences by parser category, with the rare categories shown on an expanded scale. Fallback marks captions from which no phrase was extracted and whose full text is used as a single concept.

Figure 4(b)–(c) shows the semantic composition of these queries. Noun phrases predominate and provide queries about entities, sound sources, and scene elements; action phrases add event processes, allowing the same audio to be read in terms of both objects and actions. Compared with the training set, the test set has a higher share of action phrases and a lower share of noun phrases, yielding a different mix of concept types. Spatial phrases occur less frequently but enter alignment as distinct concepts when extracted, bringing positional relationships into the conditioned representations. When no phrase is extracted, the parser retains the full caption as a single concept. Such a caption still participates in audio–concept alignment and audio recomposition, but contributes no

within-audio concept pair to the separation objective. Single- and multi-concept captions thus follow the same training procedure, with their concept sets determining the corresponding alignment, soft-positive, and separation relationships.

C Momentum Queue Hyperparameters

With concept-level contrast fixed to C2A-only, we vary the queue capacity and the number of selected negatives to select the main configuration. Table 7 reports inference results on AudioCaps.

Table 7: AudioCaps momentum-queue capacity and negative selection. M denotes the FIFO queue capacity (stored keys per queue) and K_{neg} the number of selected negatives. Capacity comparisons retain Top-24 and negative-count comparisons retain capacity 2400; concept-level momentum contrast is C2A-only in both blocks. Values are percentages (mean ± sample standard deviation).

Setting	**A2T**	**T2A**
	R@1	R@1
$M = 1200$	53.64 ± 0.42	43.27 ± 0.42
$M = 2400$	55.49 ± 0.10	43.01 ± 0.24
$M = 4800$	54.44 ± 1.16	43.21 ± 0.52
$K_{\text{neg}} = 12$	54.06 ± 0.40	43.14 ± 0.43
$K_{\text{neg}} = 24$	55.49 ± 0.10	43.01 ± 0.24
$K_{\text{neg}} = 48$	54.16 ± 0.12	43.31 ± 0.11

Queue Capacity. With Top-24 fixed, changing the queue capacity does not change the number of negatives entering each contrastive loss; it changes the historical candidate pool from which those negatives are selected. A capacity of 1200 offers a narrower ranking pool and less coverage of competitors from different batches. Expanding the queue brings more historical keys into the ranking, allowing additional near-query candidates to enter the hard-negative set. Further expansion also extends the time keys remain in the FIFO queue, changing the balance between recently produced and older representations in the candidate pool. Queue capacity thus changes the composition of the fixed Top-24 window, rather than simply adding negatives to the loss. A2T varies more across capacities than T2A in Table 7; we retain 2400 as the evaluated balance between cross-batch candidate coverage and queue turnover.

Negative Selection. At a fixed queue capacity, K_{neg} directly determines which ranks enter the contrastive loss. Top-12 concentrates supervision on the few nearest candidates but excludes competitors at subsequent ranks that still provide discrimination. Top-48 additionally includes candidates ranked 25–48, whose lower similarities add terms to the softmax denominator. Because all selected negatives share that normalization, adding these candidates redistributes negative probability mass and reduces the share assigned to the highest-ranked competitors. Top-24 retains a broader range of close competitors than Top-12 while keeping supervision concentrated on hard candidates. In concept-level C2A contrast, this window distinguishes online concept queries from queued Conditioned Audio Tokens; global momentum contrast continues to operate in both audio-to-text and text-to-audio directions.

D Stage-Level Efficiency Comparison

SAYAKA and CLAP use the same retrieval procedure: audio and text are independently encoded into global embeddings, and candidates are scored by cosine similarity between those embeddings. Although SAYAKA uses concept-conditioned representations, a recomposer, and momentum queues during training, none of these components needs to run for each audio–text candidate pair at inference. In particular, inference requires neither concept-wise scoring nor comparisons between individual concepts and audio segments. The added training objectives therefore leave the computation in the matching stage unchanged; on both datasets, SAYAKA matches CLAP's measured matching-stage and ranking-stage times. Its advantage is not faster inference than CLAP, but higher retrieval accuracy at the same measured matching and ranking cost.

Table 8: R@1 sums (%), matching-plus-ranking times (ms), CLAP-relative stage and training-step ratios, and peak memory. Stage times for CLAP, ML-ACT, MGA-CLAP, FLAM, SupCLAP, Cacophony, and SAYAKA use updated measurements; m-LTM, GPA, and THA+DCR retain reported totals. Training resources come from independent short tests.

Dataset	Method	R@1 sum	M+R (ms)	M+R (× CLAP)	Peak MiB	Train (× CLAP)
AudioCaps	CLAP	86.26	1.264	1.000	1034	1.00
	ML-ACT	86.80	1.264	1.000	1034	1.03
	THA+DCR	93.68	177,086.93	140,100.419	5075	2.24
	GPA	89.13	287.80	227.690	1063	1.05
	MGA-CLAP	92.52	1.264	1.000	1065	1.04
	FLAM	87.77	1.264	1.000	1060	1.03
	m-LTM	93.65	4.94	3.908	1038	3.53
	SupCLAP	88.79	1.264	1.000	1034	1.07
	Cacophony	88.63	1.264	1.000	1273	1.36
	SAYAKA	**98.50**	1.264	1.000	1974	1.68
Clotho	CLAP	35.83	1.668	1.000	1592	1.00
	ML-ACT	35.42	1.668	1.000	1592	1.00
	THA+DCR	35.16	222,803.56	133,575.276	5690	1.22
	GPA	37.13	367.88	220.552	1621	1.01
	MGA-CLAP	39.43	1.668	1.000	1622	1.02
	FLAM	36.00	1.668	1.000	1618	1.01
	m-LTM	39.39	5.50	3.297	1596	1.48
	SupCLAP	37.40	1.668	1.000	1592	1.02
	Cacophony	35.97	1.668	1.000	1831	1.07
	SAYAKA	**40.02**	1.668	1.000	2532	1.18

The higher stage times of other methods arise chiefly in the matching stage rather than the ranking stage. When candidate scoring extends beyond a single global-vector similarity to comparisons involving local representations, prototypes, or more elaborate cross-modal relationships, the additional work becomes part of candidate matching and is repeated across candidates. Table 3 shows this distinction for m-LTM, GPA, and THA+DCR: their ranking-stage times remain close to CLAP's, whereas their matching-stage times are higher. The difference in total time thus lies primarily in computing candidate scores, not in sorting those scores afterward.

This design places the additional computation in training. SAYAKA uses more peak memory and takes longer per training step than CLAP, but its concept-conditioned branch, recomposer, and momentum queues are absent from the retrieval path. Several other methods also match CLAP's measured stage time; what distinguishes SAYAKA in Table 8 is the highest R@1 sum on both datasets at that same matching-plus-ranking time.