Omni-Embed-Audio: Leveraging Multimodal LLMs for Robust Audio-Text Retrieval

HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo, and Du-Seong Chang
Sogang University
{judejiwoo, sysky309, dlstjd6474, mwkoo, dschang}@sogang.ac.kr

Abstract

Audio-text retrieval systems based on Contrastive Language-Audio Pretraining (CLAP) achieve strong performance on traditional benchmarks; however, these benchmarks rely on caption-style queries that differ substantially from real-world search behavior, limiting their assessment of practical retrieval robustness. We present Omni-Embed-Audio (OEA), a retrieval-oriented encoder leveraging multimodal LLMs with native audio understanding. To systematically evaluate robustness beyond caption-style queries, we introduce User-Intent Queries (UIQs)—five formulations reflecting natural search behaviors: questions, commands, keyword tags, paraphrases, and exclusion-based negative queries. For negative queries, we develop a hard negative mining pipeline and propose discrimination metrics (HNSR, TFR) assessing models' ability to suppress acoustically similar distractors. Experiments on AudioCaps, Clotho, and MECAT show that OEA achieves comparable text-to-audio retrieval performance to state-of-the-art M2D-CLAP, while demonstrating clear advantages in two critical areas: (1) **dominant text-to-text retrieval** (+22% relative improvement), and (2) **substantially superior hard negative discrimination** (+4.3%p HNSR@10, +34.7% relative TFR@10), revealing that LLM backbones provide superior semantic understanding of complex queries.

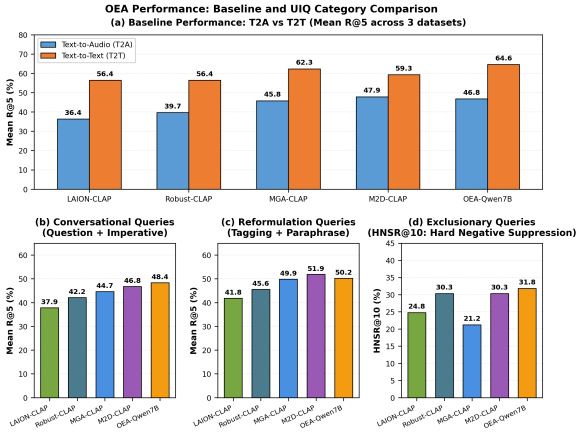

Figure 1: Performance comparison using OEA-Qwen7B (+Cl) as representative model (mean R@5 across AudioCaps, Clotho, and MECAT). (a) Baseline performance: While M2D-CLAP leads T2A (47.9%), OEA achieves competitive results (46.4%) and substantially outperforms all baselines on T2T (64.8% vs. M2D-CLAP 59.3%, +5.5). (b–d) UIQ analysis: M2D-CLAP shows strong UIQ retrieval; however, OEA achieves best Imperative query performance (49.9% vs. 44.7%) and substantially outperforms on hard negative discrimination metrics (HNSR@10: 34.6% vs. 30.3%, +4.3), demonstrating superior semantic understanding of complex queries.

1 Introduction

Text-based audio retrieval has emerged as a critical capability for navigating the exponentially growing repositories of audio content across multimedia production, gaming, filmmaking, and creative industries (Font et al., 2013; Koepke et al., 2023). Open-vocabulary audio-language models (ALMs), particularly those based on Contrastive Language-Audio Pretraining (CLAP) (Niizumi et al., 2025; Li et al., 2024; Wu et al., 2023; Elizalde et al., 2023), have achieved remarkable success by learning joint embedding spaces that align audio and text representations. These models enable users to retrieve relevant audio using natural language queries, offering an intuitive interface for searching through vast audio databases.

Despite significant progress on standard benchmarks, recent studies have revealed fundamental limitations in how audio retrieval systems are evaluated. Standard benchmarks such as AudioCaps and Clotho rely on caption-style queries that closely mirror training data distributions, creating an evaluation paradigm that may not reflect real-world retrieval performance. Selvakumar et al. (2024) demonstrated that existing ALMs suffer performance degradations of up to 16% when faced

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22664–22680

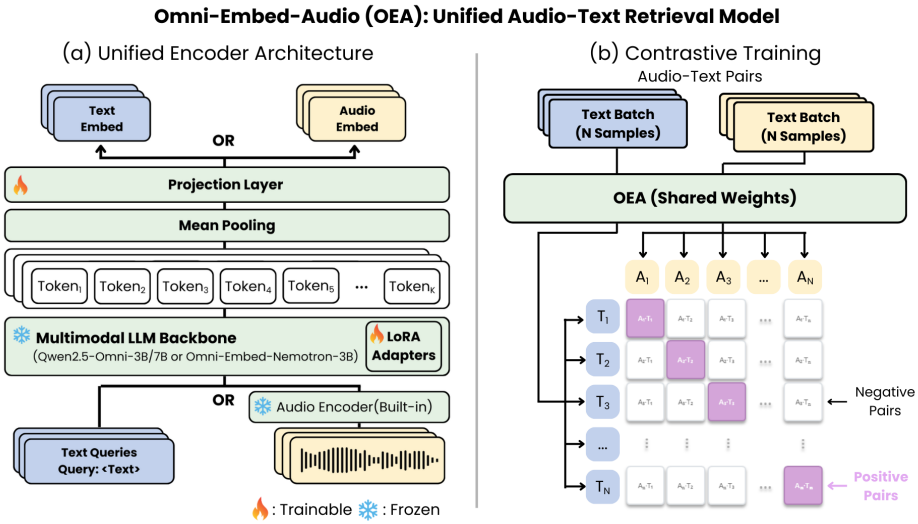

Figure 2: OEA architecture overview. A shared multimodal LLM backbone processes both text and audio inputs. LoRA adapters are applied to attention layers, and modality-specific projection heads map representations to a shared 512-dimensional embedding space with L2 normalization.

with paraphrased queries, while Weck and Font (2024) found substantial mismatch between real-world query patterns (averaging 1.8 tokens) and the descriptive captions used for training. These findings expose a critical gap: current benchmarks may overestimate model capabilities because they test on caption-style queries rather than the diverse *query-type variations* characterizing real-world usage.

To address these limitations, we present **Omni-Embed-Audio (OEA)**, a retrieval-oriented multimodal encoder that leverages pre-trained LLMs with native audio understanding. Unlike traditional dual-encoder approaches, OEA employs a single shared transformer backbone for both modalities, combined with parameter-efficient LoRA adaptation. We hypothesize that this LLM-based approach can benefit from superior language understanding capabilities, potentially yielding improved performance even on traditional text-to-audio retrieval benchmarks.

To systematically evaluate retrieval robustness, we introduce **User-Intent Queries (UIQs)**—five query formulations organized into three categories reflecting distinct aspects of real-world search behavior: (1) *Conversational Queries* (questions, commands) anticipating the shift toward speech-based AI interfaces; (2) *Reformulation Queries* (tags, paraphrases) testing robustness to query variation; and (3) *Exclusionary Queries* (negative queries) assessing fine-grained semantic discrimination.

Extensive experiments demonstrate that while recent CLAP models like M2D-CLAP achieve strong text-to-audio retrieval performance, OEA provides complementary strengths: (1) dominant text-to-text retrieval performance (+5.5%p over M2D-CLAP), (2) superior performance on imperative queries (+5.1%p), and critically, (3) substantially improved hard negative discrimination (+4.3%p HNSR@10). These discrimination improvements emerge without explicit UIQ training, suggesting LLM backbones provide superior semantic understanding of complex queries involving negation and fine-grained distinctions.

Our contributions are:

1. We present **OEA**, a unified encoder architecture leveraging multimodal LLMs, demonstrating dominant T2T performance and superior semantic understanding of complex queries.

2. We identify **query-type robustness** as a critical dimension of audio retrieval evaluation and introduce the **UIQ benchmark** with five query types organized into three linguistically-motivated categories.

3. We propose **novel hard negative-based evaluation metrics** (HNSR, TFR, Δ-Rank) that reveal limitations in standard retrieval metrics and demonstrate OEA's superior discrimination capabilities.

To facilitate future research, we release three UIQ benchmark datasets (**AudioCaps-UIQ**, **Clotho-UIQ**, and **MECAT-UIQ**) along with an interactive web demo for exploring model behavior across query types.[1]

2 Related Work

2.1 Audio-Text Retrieval

Contrastive Language-Audio Pretraining (CLAP) has become the dominant paradigm for audio-text retrieval (Li et al., 2024; Wu et al., 2023; Elizalde et al., 2023). LAION-CLAP (Wu et al., 2023) scaled training to 630K audio-text pairs with feature fusion and keyword-to-caption augmentation. MGA-CLAP (Li et al., 2024) introduced multi-granularity aggregation for improved alignment, while CompA (Ghosh et al., 2024) addressed compositional reasoning through composition-aware hard negatives. Most recently, M2D-CLAP (Niizumi et al., 2025) combines self-supervised masked modeling (M2D) with CLAP, achieving state-of-the-art text-to-audio retrieval performance by learning general-purpose audio-language representations that excel in both zero-shot and transfer learning scenarios. However, these models are trained and evaluated primarily on caption-style queries, leaving their robustness to diverse query formulations unexplored.

2.2 Real-World Query Patterns and Benchmark Limitations

Standard benchmarks such as AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) rely on caption-style queries that mirror training distributions. However, Weck and Font (2024) found that real Freesound queries average only 1.8 tokens, and Penha et al. (2022) demonstrated ∼20% effectiveness drops with paraphrased queries, revealing that single-format benchmarks mask significant model weaknesses.

2.3 Query Robustness and Negation in Retrieval

Query sensitivity is well-documented: Selvakumar et al. (2024) proposed RobustCLAP achieving 0.8–13% improvements on paraphrases. For negation, NevIR (Weller et al., 2023) and ExcluIR (Zhang et al., 2024) showed that neural retrievers struggle with exclusionary queries, with performance barely above random. These findings motivate our UIQ benchmark and hard negative discrimination metrics.

3 Methodology

3.1 OEA Architecture

Omni-Embed-Audio (OEA) is a retrieval-oriented multimodal encoder designed for audio-text retrieval. The key insight is to leverage pre-trained large language models (LLMs) with native audio understanding capabilities as a unified encoder for both text and audio modalities.

Unified Encoder Architecture Unlike traditional dual-encoder approaches that use separate encoders for text and audio, OEA uses a **single shared transformer backbone** for both modalities (Figure 2). Text and audio inputs are processed through the same transformer, reducing the modality gap and enabling audio understanding to benefit from the LLM's rich language priors.

Input Processing For text encoding, queries are wrapped with a `query:` prefix, tokenized, passed through the transformer, and mean-pooled over the last hidden layer. For audio encoding, raw waveforms (16kHz mono) are processed through the model's native audio encoder with a `passage:` prefix and identical mean pooling, producing modality-agnostic representations.

Parameter-Efficient Adaptation We employ LoRA adaptation with lightweight adapter matrices attached to attention layers, along with modality-specific projection heads that compress backbone representations into a shared 512-dimensional retrieval embedding space. All backbone weights remain **frozen**; only LoRA adapters and projection heads are trained, yielding approximately 11–16M trainable parameters.

Training Objective We use symmetric contrastive learning with InfoNCE loss ($\tau = 0.07$). Implementation details are provided in Appendix A.

Backbone Models We instantiate OEA with three backbone scales: Omni-Embed-Nemotron-3B (Xu et al., 2025b) (∼3B parameters), Qwen2.5-Omni-3B (Xu et al., 2025a) (∼3B parameters), and Qwen2.5-Omni-7B (Xu et al., 2025a) (∼7B parameters). All backbones are multimodal LLMs with native audio understanding.

[1] UIQ benchmarks: 13,053 queries total (AudioCaps: 4,530, Clotho: 4,722, MECAT: 3,801). Web demo includes 75 representative samples (15 per query type): https://omni-embed-audio.github.io

3.2 User-Intent Queries (UIQs)

3.2.1 UIQ Taxonomy

To address the gap between benchmark queries and real-world search behavior identified in Section 2.2, we introduce **User-Intent Queries (UIQs)**: five query formulations organized into three categories reflecting distinct aspects of real-world search behavior.

Category 1: Conversational Queries These query types anticipate the shift from text-based to speech-based AI interfaces, where users interact through natural dialogue rather than keyword input.

- **Question Query**: Natural language questions common in conversational search and voice assistants (e.g., *"Can you find clear dog barks echoing in a large hall?"*).
- **Imperative Query**: Direct command-style queries reflecting how users interact with AI assistants (e.g., *"Find crisp footsteps on gravel with light echo"*). LLM backbones pretrained on instruction-following data should excel at parsing such commands.

Category 2: Reformulation Queries These query types test robustness to the natural variation in how users express identical information needs, addressing the vocabulary mismatch problem central to information retrieval (Carpineto and Romano, 2012).

- **Keyphrase Query**: Keyword-style queries with comma-separated tags preferred by users seeking concise searches (e.g., *"dog barks, echoing hall, reverberant"*). This format reflects actual Freesound query patterns and tests whether models rely on syntactic structure.
- **Paraphrase Query**: Query reformulation has been shown to significantly improve retrieval performance in information retrieval systems (Ma et al., 2023; Jang et al., 2024). Motivated by this, we include paraphrase queries—declarative descriptions that rephrase audio content using varied vocabulary and structure (e.g., *"Echoing dog barks resonate through a large empty hall"*)—to evaluate whether retrieval models maintain robustness when queries are reformulated while preserving semantic intent.

Query Type	Human		LLM	
	Mean	Std	Mean	Std
Question	4.26	0.91	4.73	0.44
Imperative	4.16	1.06	4.33	0.87
Keyphrase	4.35	1.05	4.47	0.50
Paraphrase	4.14	1.21	4.67	0.47
Negative	3.82	1.25	3.93	0.57
Overall	**4.15**	1.12	**4.43**	0.66

Table 1: UIQ benchmark validation comparing human evaluation (9 annotators, 675 ratings) with LLM evaluation (Claude Opus 4.5).

Category 3: Exclusionary Queries This category tests fine-grained semantic discrimination through queries specifying both desired and excluded content (see Figure 3 for an illustrative example).

- **Negative Query**: Queries specifying desired content AND explicit exclusions (e.g., *"Heavy rain and wind on metal surfaces without thunder or engine noise"*). Each negative query is grounded in a pre-mined (target audio, hard negative audio) pair, enabling quantitative evaluation of exclusion understanding.

3.2.2 UIQ Generation Process

For positive query types (Question, Imperative, Paraphrase, Keyphrase), we generate queries using GPT-5.1 with vocabulary grounding constraints requiring queries to reuse wording from original captions. To avoid length-induced distribution shift in evaluation, we maintain query lengths within ± 2 words of original captions, preserving semantic completeness for controlled evaluation. For negative queries, we additionally provide hard negative captions to craft precise exclusion statements. Human reviewers verified sampled queries and flagged semantically incorrect outputs for regeneration, ensuring benchmark quality.

3.2.3 UIQ Benchmark Validation

We construct three evaluation datasets: **AudioCaps-UIQ**, **Clotho-UIQ**, and **MECAT-UIQ**, releasing sample subsets for reproducibility and the full datasets upon publication. We validate these benchmarks through both LLM-based evaluation (Claude Opus 4.5) and human annotation, using the same evaluation criterion: *"Does this make sense as a search query regarding the original captions?"* Responses are rated on a 5-point Likert scale (see Appendix C for scale definitions).

To assess whether UIQs reflect practical search behavior beyond synthetic generation, we additionally compare their token-length distribution with real Freesound queries in Appendix I.

Evaluation Methodology We validate UIQ quality through two complementary approaches: (1) **Human evaluation**: 9 annotators rated 75 UIQ samples (15 audio clips $\times$ 5 UIQ types) on a 5-point Likert scale, yielding 675 total ratings. Annotators were recruited through an internal Sogang University community posting open to undergraduate and graduate students, and were paid an hourly wage; (2) **LLM-based evaluation**: Claude Opus 4.5 evaluated the same samples using identical criteria.

Validation Results Table 1 presents validation results. Human evaluation shows strong validity across all UIQ types (mean 4.15/5.0), with Question and Keyphrase queries receiving highest ratings. LLM evaluation demonstrates similar patterns with moderate Human-LLM agreement. Imperative queries show strongest agreement ($r = 0.89$, $p < 0.001$), confirming the quality of generated UIQs.

3.3 Hard Negative Mining and Evaluation Metrics

Evaluating exclusionary query understanding requires paired examples where models must distinguish acoustically similar but semantically distinct audio. We develop a four-stage hard negative mining pipeline that combines acoustic similarity retrieval (using MGA-CLAP embeddings) with semantic dissimilarity filtering (using BGE sentence embeddings), yielding audio pairs that are acoustically confusable but semantically distinct. Candidate pairs were additionally reviewed by human annotators before finalization to verify that they were genuinely confusing from an auditory perception standpoint rather than artifacts of a single embedding model. Details are provided in Appendix K.

Standard retrieval metrics (R@k) assess whether models retrieve relevant audio but do not measure whether models correctly suppress confusable alternatives. Figure 3 illustrates this distinction. We introduce **HNSR@k** (Hard Negative Suppression Rate at k): the percentage of queries where the target is retrieved within top-k AND the hard negative is ranked outside top-k. This metric captures both successful retrieval and distractor suppression—the core challenge of exclusionary queries. We also report **Δ-Rank** = Rank(HN) $-$ Rank(T) as a continuous measure of target-distractor separation, where higher values indicate the model more effectively pushes hard negatives away from targets in the ranking. Additional metrics (TFR, TFR-HN@k) are detailed in Appendix L.

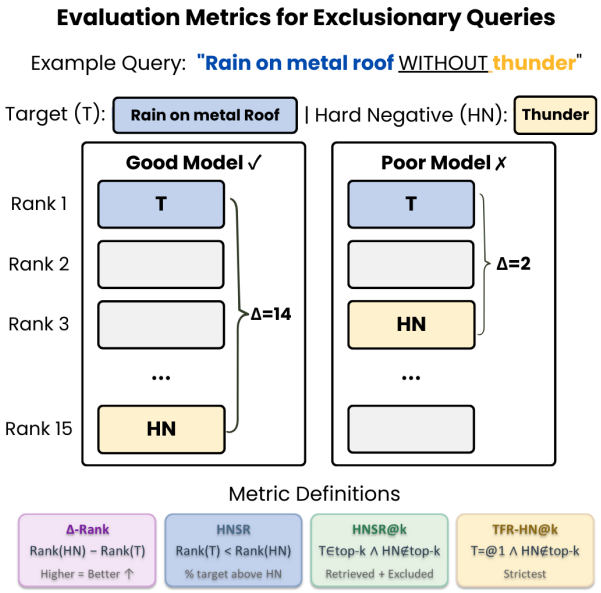

Figure 3: Evaluation metrics for exclusionary queries. Standard R@k metrics only check if the Target audio is retrieved, but do not verify whether the model suppresses the Hard Negative (acoustically similar distractor). Our proposed HNSR@k explicitly measures the model's ability to rank the Target above the Hard Negative.

4 Experiments

4.1 Datasets

Training Data OEA training follows a multi-stage curriculum: WavCaps (275,618 samples filtered to $\leq$31 seconds) for initial audio-text alignment, then AudioCaps v2 (91,256 training samples) for caption-based retrieval. Optionally, additional Clotho v2 training (3,839 clips) improves performance on natural audio descriptions. Models trained without Clotho are denoted as **OEA**; those with additional Clotho training are denoted **OEA (+Cl)**.

Evaluation Data We evaluate on three datasets: **AudioCaps** (975 clips, 5 captions each), **Clotho** (1,045 clips, 5 captions each), and **MECAT** (847 auto-captioned pairs).

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLAP Models									
LAION-CLAP	22.15	54.63	69.93	13.34	33.49	46.22	6.80	20.93	30.66
Robust-CLAP	26.48	60.12	75.53	14.83	37.89	50.05	6.51	21.05	31.39
MGA-CLAP	36.39	70.11	82.46	18.78	43.79	56.63	**8.02**	23.51	34.08
M2D-CLAP	**41.39**	**77.13**	**88.27**	17.55	42.91	55.54	7.37	23.55	33.63
Vanilla LALMs (No Retrieval Training)									
Nemotron-3B	3.73	9.09	13.11	7.20	21.57	30.12	4.09	13.27	21.99
Qwen2.5-Omni-3B	0.37	1.19	1.87	0.17	0.63	1.13	0.12	0.63	1.57
Qwen2.5-Omni-7B	0.41	1.29	2.34	0.10	0.65	1.32	0.16	0.77	2.38
OEA Models (Ours)									
OEA-Nemo3B	38.87	72.64	83.88	19.04	40.57	54.24	7.96	**24.53**	**35.81**
OEA-Nemo3B (+Cl)	35.75	68.64	81.35	21.57	47.16	60.36	6.98	23.00	33.41
OEA-Qwen3B	37.60	71.16	83.73	19.18	42.05	55.85	5.86	17.96	25.94
OEA-Qwen3B (+Cl)	35.96	69.35	81.99	**22.87**	**49.78**	**63.25**	5.78	17.16	24.76
OEA-Qwen7B	38.03	72.25	84.53	19.77	44.78	57.65	7.74	23.29	33.49
OEA-Qwen7B (+Cl)	34.28	67.47	80.84	22.53	48.80	62.70	7.02	22.94	33.12

Table 2: Text-to-audio (T2A) baseline results (%). M2D-CLAP achieves strong performance on AudioCaps, while OEA models excel on Clotho and MECAT. **Bold**: best per column, underline: second best.

A critical consideration for audio retrieval evaluation is **data leakage** between training and evaluation sets. WavCaps aggregates audio from multiple sources including AudioSet Strongly-Labeled, Freesound, BBC Sound Effects, and SoundBible. This creates potential overlap with evaluation benchmarks: AudioCaps derives from AudioSet, while Clotho sources from Freesound. Our analysis (Appendix B) confirms significant overlap—17.7% of AudioCaps test clips and 61.0% of Clotho evaluation clips appear in WavCaps subsets.

To provide a **leakage-free evaluation**, we additionally evaluate on **MECAT** (Wu et al., 2025), which derives audio from ACAV100M (Lee et al., 2021)—a large-scale audio-visual dataset sourced from web videos that has no overlap with WavCaps training sources. While MECAT uses auto-generated captions (potentially noisier than human annotations in AudioCaps and Clotho), it provides an uncontaminated benchmark for assessing true generalization performance. Details of our leakage analysis are provided in Appendix B.

4.2 Models and Training

Baselines We compare against four CLAP variants: LAION-CLAP (Wu et al., 2023), Robust-CLAP (Selvakumar et al., 2024), MGA-CLAP (Li et al., 2024), and M2D-CLAP (Niizumi et al., 2025)—the current state-of-the-art CLAP model that combines self-supervised masked modeling with contrastive language-audio pretraining. To isolate the contribution of our training approach, we also evaluate **vanilla LALMs**—Qwen2.5-Omni (3B and 7B) and Nemotron-3B without any retrieval-specific training. These baselines use mean-pooled last hidden states with L2 normalization, identical to OEA's embedding extraction.

OEA Variants We train six OEA variants across three backbones (Nemotron-3B, Qwen3B, Qwen7B): base OEA models and OEA (+Cl) variants with additional Clotho training.

Training Configuration Training uses symmetric InfoNCE loss with $\tau = 0.07$, AdamW optimizer (lr 3e-4 or 5e-4), and PyTorch DDP with BFloat16 precision. Training proceeds in stages with early stopping based on validation R@10.

5 Results

5.1 Text-to-Audio Retrieval

Table 2 presents text-to-audio (T2A) retrieval results. Without retrieval-specific training, vanilla LALMs achieve only $\sim$1% R@5, confirming that adaptation is essential for cross-modal retrieval.

On AudioCaps, M2D-CLAP achieves the strongest T2A performance (77.13% R@5), while OEA models show competitive results (up to 72.64% R@5). On the remaining benchmarks, OEA achieves comparable or superior performance: on Clotho, OEA-Qwen3B (+Cl) achieves best R@10 (63.25%, +7.71 over M2D-CLAP), and on MECAT (leakage-free), OEA-Nemo3B achieves

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLAP Models									
LAION-CLAP	41.11	63.94	74.19	50.05	65.88	72.67	18.06	39.35	50.57
Robust-CLAP	41.09	63.96	74.22	50.05	65.88	72.67	18.06	39.35	50.57
MGA-CLAP	47.73	71.63	80.96	63.27	74.70	78.79	19.56	40.64	52.14
M2D-CLAP	47.51	70.03	79.96	55.85	69.05	74.76	17.92	38.74	49.51
Vanilla LALMs (No Retrieval Training)									
Nemotron-3B	33.17	52.27	62.13	57.84	69.36	74.26	6.86	20.77	29.56
Qwen2.5-Omni-3B	22.63	34.48	41.39	38.35	51.14	57.07	3.18	9.75	15.39
Qwen2.5-Omni-7B	24.06	37.19	44.62	40.23	54.26	60.27	3.16	10.46	15.92
OEA Models (Ours)									
OEA-Nemo3B	48.94	**72.62**	**82.19**	62.79	74.12	78.91	22.11	43.34	54.99
OEA-Nemo3B (+Cl)	48.94	72.21	81.64	63.77	**75.29**	**80.11**	22.80	44.65	55.66
OEA-Qwen3B	49.42	72.27	81.76	62.81	73.76	78.22	22.60	45.38	56.60
OEA-Qwen3B (+Cl)	49.64	71.94	81.01	**64.52**	75.25	79.71	24.65	46.23	58.24
OEA-Qwen7B	**50.30**	72.31	82.05	63.66	74.32	79.37	25.43	47.19	58.18
OEA-Qwen7B (+Cl)	49.93	71.79	81.05	63.58	75.04	79.90	**25.45**	**47.41**	**58.67**

Table 3: Text-to-text (T2T) baseline results (%). OEA models substantially outperform all CLAP baselines including M2D-CLAP across all datasets. **Bold**: best per column, underline: second best.

best R@5/R@10 (24.53%/35.81%). Overall, **T2A performance is similar across methods**, with M2D-CLAP excelling on in-domain AudioCaps and OEA demonstrating stronger cross-domain generalization.

5.2 Caption-Based Text-to-Text Retrieval

We evaluate **text-to-text (T2T)** retrieval (Table 3), simulating caption-based audio retrieval where audio is first converted to text via automatic captioning (Chu et al., 2024; Xu et al., 2025a). T2T achieves substantially higher performance than T2A (e.g., 47.41% vs. 22.94% R@5 on MECAT), suggesting caption-based pipelines offer a practical alternative.

OEA Dominates T2T. OEA models substantially outperform all CLAP baselines: +8.67 R@1 over M2D-CLAP on Clotho (64.52% vs. 55.85%) and +22% relative improvement on MECAT R@5 (47.41% vs. 38.74%). This dominance stems from OEA's unified LLM backbone processing both query and caption text, enabling richer semantic matching compared to CLAP's separate and smaller text encoders. While CLAP models use lightweight text encoders optimized for contrastive alignment, OEA leverages the full representational capacity of billion-parameter LLMs for text understanding. This suggests that for caption-based retrieval pipelines—increasingly practical with modern audio captioning systems—LLM-based encoders offer substantial advantages over traditional CLAP architectures.

5.3 User-Intent Query to Audio Retrieval

5.3.1 Conversational and Reformulation Queries

Table 4 presents UIQ performance across Question, Imperative, Keyphrase, and Paraphrase query types.

UIQ Performance: Comparable with Complementary Strengths For standard UIQ retrieval, M2D-CLAP and OEA show **comparable overall performance** with complementary strengths. M2D-CLAP achieves the highest mean scores on Question (48.76%), Keyphrase (53.16%), and Paraphrase (50.58%) queries, while OEA models achieve the best performance on **Imperative** queries—the most command-like formulation. OEA-Qwen7B (+Cl) achieves 49.87% compared to M2D-CLAP's 44.74%—a +5.13 point gain (+11.5% relative), confirming that LLM backbones excel at parsing command-style queries. M2D-CLAP achieves the best overall **Avg UIQ** (47.76%) with OEA-Qwen3B (+Cl) achieving competitive second place (47.18%, -0.58%p). This split is consistent with the underlying training signals: M2D-CLAP's contrastive pretraining is optimized over descriptive captions, which align closely with Question, Keyphrase, and Paraphrase formulations, whereas OEA benefits from instruction-tuned LLM backbones that parse directive verbs ("find", "retrieve") as goal-conditioned commands rather than

Model	Conversational		Reformulation		Exclusionary	Avg UIQ
	Question	Imperative	Keyphrase	Paraphrase	Hard Neg.	
LAION-CLAP	38.65	37.06	42.71	40.92	24.7	38.01
Robust-CLAP	43.24	41.07	47.19	43.98	30.3	42.07
MGA-CLAP	44.41	44.94	50.97	48.74	21.2	45.16
M2D-CLAP	**48.76**	44.74	**53.16**	**50.58**	30.3	**47.76**
OEA-Nemo3B	46.65	47.58	48.87	49.19	32.9	46.39
OEA-Nemo3B (+Cl)	46.10	49.41	49.02	48.82	32.5	46.97
OEA-Qwen3B	44.17	46.06	48.20	46.94	31.6	44.57
OEA-Qwen3B (+Cl)	46.71	48.55	50.07	49.05	32.8	47.18
OEA-Qwen7B	47.47	49.30	50.31	50.12	31.8	47.08
OEA-Qwen7B (+Cl)	47.32	**49.87**	50.21	50.25	**34.6**	47.16

Table 4: Mean UIQ results across AudioCaps, Clotho, and MECAT (%). Queries are organized by category: Conversational (Question, Imperative), Reformulation (Keyphrase, Paraphrase), and Exclusionary (Hard Neg.). Conversational, Reformulation, and Avg UIQ use R@5; Exclusionary uses HNSR@10 to measure hard negative discrimination ability. M2D-CLAP achieves best Avg UIQ, while OEA models achieve best Imperative and Hard Negative discrimination. **Bold**: best, underline: second best.

surface tokens. Notably, OEA's key advantages emerge in **T2T retrieval** (+22% improvement) and **hard negative discrimination** (detailed below), rather than standard UIQ performance.

5.3.2 Exclusionary Query Evaluation

Negative queries represent the most challenging UIQ type, requiring models to understand both desired content and explicit exclusions. As shown in Table 4, the Exclusionary category reports HNSR@10 to measure hard negative discrimination ability.

Standard Retrieval vs. Discrimination M2D-CLAP achieves the best standard retrieval on exclusionary queries (41.56% R@5), contributing to its highest overall Avg UIQ. However, **standard metrics do not capture discrimination ability**—the core challenge where models must distinguish between acoustically similar but semantically distinct audio.

OEA Dominates Hard Negative Discrimination OEA-Qwen7B (+Cl) achieves the best HNSR@10 (34.6% vs. M2D-CLAP's 30.3%, +4.3%p), demonstrating superior hard negative discrimination. Additional metrics in Appendix M confirm this pattern: OEA achieves HNSR 74.4% vs. M2D-CLAP's 68.0% (+6.4%p), and TFR@10 10.1% vs. 7.5% (+34.7% relative improvement).

Implications These results reveal a critical distinction: while T2A retrieval and standard UIQ performance show **comparable results between M2D-CLAP and OEA**, the key differentiators emerge in two areas where OEA demonstrates clear superiority: (1) **T2T retrieval**, where OEA achieves +22% relative improvement leveraging its LLM backbone for richer text-to-text semantic matching, and (2) **hard negative discrimination**, where OEA substantially outperforms (+4.3%p HNSR@10, +34.7% relative TFR@10). For exclusionary queries where users explicitly specify what they *don't* want, discrimination ability is arguably more important than raw retrieval performance. We hypothesize these advantages stem from LLM backbones' exposure to negation patterns ("do not", "without", "except") during instruction-following pretraining and their superior semantic understanding of nuanced language. Unlike CLAP text encoders, which compress the entire query into a single bag-of-content vector, LLM-based encoders preserve compositional structure through attention, making it easier to propagate exclusion cues rather than absorb them as additional positive keywords—a failure mode consistent with prior negation studies in text retrieval (Weller et al., 2023; Zhang et al., 2024).

5.4 Efficiency and Backbone Generality

Table 5 summarizes efficiency. Although OEA uses larger backbones than CLAP baselines, its online retrieval latency remains practical because audio embeddings are pre-computed offline and per-query serving depends only on text encoding. OEA-Nemo3B encodes text in 2.3 ms/query, while LoRA updates only 13.7M parameters (0.29% of the full model). Additional deployment analysis in Appendix H shows that the latency gap between OEA-Nemo3B and OEA-Qwen3B is driven more by architectural overhead than nominal model size.

Our multi-backbone results also clarify general-

Model	Audio ms/clip	Text ms/query	GPU (GB)	Params (M)
LAION-CLAP	107.7	0.53	∼0.6	158
MGA-CLAP	31.7	0.72	∼0.6	148
M2D-CLAP	58.1	0.30	0.7	89
OEA-Nemo3B	163.8	2.30	11.5	13.7
OEA-Qwen3B	539.3	2.60	11.6	16.2
OEA-Qwen7B	666.8	4.86	18.3	17.2

Table 5: Inference efficiency on Clotho (1,045 clips; A100-SXM4-80GB). Audio encoding is offline and amortized; online serving depends on text encoding latency. GPU: peak inference memory; Params: trainable (M). OEA fine-tunes only 0.29–0.36% of total parameters via LoRA.

ity. Across Nemotron-3B, Qwen2.5-Omni-3B, and Qwen2.5-Omni-7B, OEA consistently improves T2T retrieval over CLAP baselines and remains competitive on T2A, indicating that the core findings are not tied to a single vendor or scale. At the same time, 7B does not uniformly outperform 3B, suggesting that retrieval quality is bottlenecked more by contrastive alignment and dataset-backbone fit than by raw parameter count; we discuss this retrieval-specific scaling behavior in Appendix J.

6 Conclusion

We presented **Omni-Embed-Audio (OEA)**, a unified encoder leveraging multimodal LLMs for audio-text retrieval. For text-to-audio (T2A) retrieval, OEA achieves **comparable performance** to M2D-CLAP overall, with stronger cross-domain generalization on Clotho and MECAT. OEA's key advantages emerge in two critical areas: (1) **Text-to-text (T2T) retrieval**, where OEA *substantially outperforms* all baselines (+8.67%p R@1 on Clotho, +22% relative improvement on MECAT), demonstrating that LLM backbones provide superior text-to-text semantic matching for caption-based retrieval pipelines; and (2) **Hard negative discrimination**, where OEA achieves +4.3%p HNSR@10 and +34.7% relative TFR@10 improvement, revealing superior semantic understanding of complex queries involving negation. We introduced the **UIQ benchmark** with five query types and novel **discrimination metrics** (HNSR, TFR), exposing that standard retrieval metrics fail to capture models' ability to distinguish acoustically similar but semantically distinct audio—a critical capability for real-world search applications. These findings suggest a practical decision rule for deployment: M2D-CLAP remains preferable when the primary workload is in-domain caption-style T2A retrieval on benchmarks aligned with its training distribution, whereas OEA is the stronger choice for caption-indexed (T2T) pipelines, exclusion-aware queries, and cross-domain settings where robustness to query reformulation matters more than peak in-domain accuracy.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents, Contribution: 50%). This research was also supported by the MSIT, Korea, under the Top-Tier AI Global HRD invitation program (RS-2025-25461932) supervised by the IITP.

Limitations

Our work has several limitations. First, OEA currently depends on multimodal LLM backbones with native audio understanding, which narrows the space of base encoders and makes direct comparison with text-only LLM backbones non-trivial. Extending the framework to hybrid designs with a separate audio front-end is an important direction for future work. Second, although our efficiency analysis shows practical online latency, OEA still requires substantially more memory than compact CLAP models, so deployment on edge hardware may require quantization or distillation. Third, our hard negative set is filtered with MGA-CLAP and BGE before human verification; while this combination reduces single-model bias, it may still miss other forms of acoustic confusion that arise in real-world search logs. Finally, UIQ generation uses a single LLM (GPT-5.1) under controlled vocabulary constraints. Our human validation and real-query distribution analysis reduce concerns about synthetic artifacts, but they do not eliminate the possibility that some query styles remain underrepresented. Future work should incorporate larger-scale human-authored search logs and organically collected exclusion queries.

References

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2024. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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

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

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 411–412. ACM.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, S Ramaneswaran, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. CompA: Addressing the gap in compositional reasoning in audio-language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. Itercqr: Iterative conversational query reformulation with retrieval guidance. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 8121–8138. Association for Computational Linguistics.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 119–132.

A Sophia Koepke, Andreea-Maria Oncescu, João F Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685.

Sangho Lee, Jiho Bui, Zihang Lu, Yuta Yoshitomo, Sung Ju Chang, Joon Son Hwang, and Gunhee Lee. 2021. ACAV100M: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10274–10284.

Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu. 2024. Advancing multi-grained alignment for contrastive language-audio pre-training. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7356–7365.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315. Association for Computational Linguistics.

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

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. 2025. M2D-CLAP: Exploring general-purpose audio-language representations beyond CLAP. *IEEE Access*.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR)*, pages 397–412. Springer.

Ramaneswaran Selvakumar, Sonal Kumar, Hemant Kumar Giri, Nishit Anand, Ashish Seth, Sreyan Ghosh, and Dinesh Manocha. 2024. Do audio-language models understand linguistic variations? *arXiv preprint arXiv:2410.16505*.

Benno Weck and Frederic Font. 2024. The language of sound search: Examining user queries in audio search engines. In *Proceedings of the 9th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2023. NevIR: Negation in neural information retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4163–4173.

Anonymous Wu and 1 others. 2025. Mecat: A benchmark for evaluating audio-text retrieval with minimal data contamination. *arXiv preprint arXiv:2507.23511*. Under review.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and

Junyang Lin. 2025a. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. 2025b. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*.

Wenhao Zhang and 1 others. 2024. ExcluIR: Exclusionary neural information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

A Implementation Details

Input Processing For text encoding, queries are wrapped in a chat template with a `query:` prefix, tokenized, passed through the transformer, and mean-pooled over the last hidden layer. For audio encoding, raw waveforms are loaded at 16kHz mono, converted to audio features via the model's native audio processor, wrapped with a `passage:` prefix, and processed through the same transformer with identical mean pooling.

Adaptation Layers LoRA is configured with rank $r = 16$, $\alpha = 32$, and dropout 0.05, targeting query, key, value, and output projections of all attention layers. Each modality has a dedicated projection head consisting of a bias-free linear layer (hidden dim $\rightarrow$ 512), dropout (0.1), layer normalization, and L2 normalization to produce unit-norm embeddings.

Training Objective We use symmetric contrastive learning with InfoNCE loss:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{t\rightarrow a} + \mathcal{L}_{a\rightarrow t}\right) \quad (1)$$

where $\mathcal{L}_{t\rightarrow a} = -\log \frac{\exp(\text{sim}(t_i,a_i)/\tau)}{\sum_j \exp(\text{sim}(t_i,a_j)/\tau)}$ and $\tau = 0.07$ is the temperature parameter.

B Data Leakage Analysis

A critical but often overlooked issue in audio retrieval evaluation is **data leakage** between training corpora and evaluation benchmarks. WavCaps (Mei et al., 2024), a widely-used pretraining dataset, aggregates audio from multiple sources that overlap with standard evaluation benchmarks. We conduct a systematic analysis of potential contamination.

B.1 Dataset Provenance

Table 6 summarizes the source relationships between training and evaluation datasets.

Evaluation Set	Original Source
AudioCaps	AudioSet
Clotho	Freesound
SoundDescs	BBC Sound Effects
FSD50K	Freesound
ESC-50	Freesound
DCASE2022	Freesound
MECAT	ACAV100M (web videos)

Table 6: Source provenance of evaluation datasets. WavCaps includes AudioSet Strongly-Labeled, Freesound, BBC Sound Effects, and SoundBible subsets, creating potential overlap with most standard benchmarks except MECAT.

B.2 AudioCaps–WavCaps Overlap

We analyzed overlap between the AudioCaps test set (4,875 caption rows covering 975 unique YouTube IDs) and the WavCaps AudioSet_SL subset (108,317 captioned clips stored as `Y<YouTubeID>.wav`). We normalized every WavCaps ID by stripping the leading 'Y' and trailing '.wav' to restore the raw 11-character YouTube ID, then matched against AudioCaps test metadata.

Findings

- **173 of 975** unique AudioCaps test clips (17.7%) are present in WavCaps AudioSet_SL (representing 0.16% of WavCaps entries)
- Because AudioCaps supplies five captions per clip, this manifests as **865** duplicated caption rows
- Example overlaps include: `6BJ455B1aAs` (rocket/missile launch with explosion), `VjSEIRnLAh8` (frying food with female speech), `ztSjcZNUY7A` (crying baby with woman speaking)
- Captions on both sides describe identical acoustic content, confirming shared underlying audio rather than accidental ID collisions

B.3 Clotho–WavCaps Overlap

We analyzed overlap between the Clotho v2 evaluation split (1,045 audio files with five captions each) and the WavCaps Freesound subset. We normalized Clotho's `file_name` field and matched it case-insensitively against filenames in the Freesound metadata JSON.

Findings

- **638 of 1,045** Clotho evaluation clips (61.0%) have identical filenames in WavCaps Freesound metadata, proving they are the exact same recordings (e.g., `Radio Garble.wav` ↔ Freesound ID 80399)
- Captions in WavCaps often mention the same acoustic events (e.g., "Simulated garbled radio traffic with crosstalk"), confirming this is not merely a naming collision
- The remaining 39% likely still originate from Freesound but were renamed during Clotho curation; additional fuzzy matching (author + duration or audio fingerprinting) would surface more overlaps

B.4 Implications and Mitigation

These findings have significant implications for audio retrieval evaluation:

- Any model trained on WavCaps AudioSet_SL without filtering overlapping IDs will have memorized nearly one-fifth of the AudioCaps test set, invalidating downstream evaluations
- Similarly, models trained on unfiltered WavCaps Freesound have already "seen" the majority of Clotho evaluation audio; reported Clotho performance is therefore inflated
- Prior experiments that mixed these datasets should be considered contaminated unless they explicitly removed overlapping samples

We apply blocklists based on our overlap analysis to exclude contaminated samples from training. To provide an uncontaminated evaluation baseline, we include **MECAT** (Wu et al., 2025), which derives from ACAV100M (Lee et al., 2021)—a dataset of web videos with no overlap with WavCaps sources. Our embedding-based and filename-based leakage checks found no significant overlap between MECAT and WavCaps. While MECAT uses auto-generated captions (potentially noisier than human annotations), it provides the only truly leakage-free benchmark among our evaluation sets.

C UIQ Benchmark Validation

To ensure UIQ quality, we validate through LLM-based evaluation (Claude Opus 4.5) and human annotation.

5-Point Likert Scale Definitions

- **1 - Incompatible**: Query meaning completely diverges from original caption or target intent
- **2 - Poor**: Query captures some elements but introduces significant semantic drift
- **3 - Acceptable**: Query roughly conveys the intended meaning with minor issues
- **4 - Good**: Query accurately reflects original meaning/intent in target format
- **5 - Excellent**: Query perfectly preserves semantics with natural formulation

Query Type	Per Dataset			Overall		
	AC	Cl	ME	Mean	Std	N
Question	4.11	4.13	4.53	4.26	0.91	135
Imperative	4.31	3.64	4.53	4.16	1.06	135
Keyphrase	4.40	4.02	4.62	4.35	1.05	135
Paraphrase	4.42	3.69	4.31	4.14	1.21	135
Negative	3.73	3.87	3.87	3.82	1.25	135
Mean	4.20	3.87	4.37	**4.15**	1.12	675

Table 7: Detailed UIQ human evaluation results by dataset. AC: AudioCaps, Cl: Clotho, ME: MECAT. 9 annotators rated 15 samples (5 per dataset) across 5 UIQ types. MECAT achieves highest validity (4.37), likely due to cleaner auto-generated captions. Clotho shows lower scores (3.87), reflecting ambiguity in some natural audio descriptions.

D UIQ Generation Prompts

This section documents the prompts used to generate User-Intent Queries (UIQs) using GPT-5.1.

D.1 System Prompt

The following system prompt instructs the model to generate all five UIQ types in a single API call:

```
You generate five retrieval queries for one
    audio example
(with a target and a negative).
Return exactly FIVE lines, one per type, in the
    form:
question: <text>
imperative: <text>
paraphrase: <text>
tagging: <text>
negative: <text>

Type-specific rules:
- question: Natural question form, 8-18 words;
  starts with Can you/Do you/Are there/Is there;
     ends with ?.
- imperative: Direct command, 8-15 words;
```

```
  starts with Find/Search for/Locate/Retrieve;
     no ?.
- paraphrase: 12-25 word declarative sentence;
  no command/question tone.
- tagging: 3-6 comma-separated tags; each 1-4
     words; lowercase.
- negative: 15-35 words; clearly state desired
     sounds AND
  exclusions using without/not/excluding.

Global rules:
- One line per type, in the exact order above.
- No numbering, bullets, or extra commentary.
```

D.2 User Prompt Template

For each audio sample, the following template is populated:

```
TARGET audio id: {target_audio_id}
NEGATIVE audio id: {negative_audio_id}

TARGET descriptions:
- {target_caption_1}
- {target_caption_2}

NEGATIVE descriptions:
- {negative_caption_1}
- {negative_caption_2}

Generate all five query types as described.
```

D.3 Generation Parameters

Parameter	Value
Model	GPT-5.1
Temperature	0.35
Top-p	0.9
Max tokens	256

Table 8: UIQ generation parameters.

D.4 Generated Query Examples

Table 9 shows example UIQs generated for a sample audio clip.

Type	Generated Query
Question	Can you find clear dog barks echoing in a large hall?
Imperative	Find crisp footsteps on gravel with light echo
Paraphrase	Echoing dog barks resonate through a large empty hall
Keyphrase	dog barks, echoing hall, reverberant
Negative	Heavy rain and wind on metal surfaces without thunder or engine noise

Table 9: Example UIQs for different query types.

E Human and LLM Evaluation Prompts

This section documents the evaluation methodology for validating UIQ quality.

E.1 Human Evaluation

Evaluation Question The web-based listening test presents annotators with the following question:

> *"Validity – Does this make sense as a search query regarding the original captions? If needed, please check the audio too."*

Rating Scale Annotators rate queries on a 5-point Likert scale:

- **1 (Poor)**: Completely invalid or nonsensical
- **2**: Major issues, barely related
- **3 (Average)**: Acceptable with some issues
- **4**: Good, appropriately reflects the caption
- **5 (Excellent)**: Perfect validity and natural formulation

Interface Each sample presents: (1) audio player for target audio, (2) original human-written captions, (3) five generated UIQs (one per type), (4) rating buttons (1–5) for each UIQ, and (5) optional comment field.

E.2 LLM Evaluation

We employ Claude Opus 4.5 for automated evaluation using the same validity criterion as human evaluation, enabling direct comparison via KL-divergence.

E.2.1 Evaluation Prompt

```
You are evaluating a generated search query for
    an audio
retrieval system.

**Original Audio Caption:** {caption}
**Generated Query:** {query}
**Query Type:** {query_type}

**Task:** Rate the validity on a scale of 1-5.

**Question:** Does this make sense as a search
    query
regarding the original caption?

**Rating Scale:**
- 1: Completely invalid or nonsensical
- 2: Poor - major issues, barely related
```

```
- 3: Acceptable - roughly conveys intent with
    issues
- 4: Good - valid, appropriately reflects
    caption
- 5: Excellent - perfectly captures intent
    naturally

Respond with ONLY: {"score": <1-5>, "reasoning":
     "<brief>"}
```

E.2.2 LLM Evaluation Parameters

Parameter	Value
Model	Claude Opus 4.5
Max tokens	256
Rate limiting	0.5s between requests

Table 10: LLM evaluation parameters.

F Detailed Baseline Results

This section provides mean results across datasets for baseline caption query evaluation.

G Detailed UIQ Results

This section provides complete per-dataset results for each UIQ type.

G.1 Question Query Results

G.2 Imperative Query Results

G.3 Paraphrase Query Results

G.4 Keyphrase Query Results

H Inference Efficiency and Deployment Cost

To complement retrieval accuracy, we benchmark inference efficiency on Clotho (1,045 clips) using a single A100-SXM4-80GB GPU. We report mean audio encoding latency, text encoding latency, peak GPU memory during inference, and the number of trainable parameters after adaptation.

Two observations are important. First, OEA remains practical for retrieval serving despite slower offline audio encoding: OEA-Nemo3B requires only 2.3 ms/query for text encoding, which is sufficient for interactive search. Second, LoRA makes adaptation parameter-efficient: only 0.29–0.36% of total parameters are updated, despite the billion-parameter backbones.

The latency gap between OEA-Nemo3B and OEA-Qwen3B (163.8 vs. 539.3 ms/clip) is not explained by backbone size alone. Both models use the same Qwen2.5-Omni audio encoder, but OEA-Nemo3B loads only the embedding-oriented thinker, whereas OEA-Qwen3B carries the full generative stack, including inactive speech-generation components that still increase memory-bandwidth pressure during inference. In deployment settings, 4-bit quantization can reduce OEA-Nemo3B memory from roughly 6 GB in bfloat16 to around 1.5 GB, making single-GPU serving feasible on consumer hardware.

I Real-World Query Distribution Analysis

A natural concern is that UIQs may overfit patterns preferred by the LLM used for generation rather than reflecting realistic user behavior. To probe this, we compare UIQ token-length statistics against the real Freesound query analysis of Weck and Font (2024), who report that user-issued sound search queries are highly concise and average 1.8 tokens.

Our UIQ design intentionally spans multiple interaction modes. Keyphrase queries are short, keyword-style formulations intended to mimic traditional search-box behavior, while Question and Imperative queries target conversational assistants. The controlled ± 2-word constraint relative to original captions prevents uncontrolled length shift across synthetic query types, allowing us to isolate query formulation rather than vocabulary novelty. This means the benchmark is not intended to exactly reproduce the full Freesound distribution; instead, it covers a broader space that includes both legacy keyword search and emerging assistant-style retrieval.

Importantly, the observed results do not match the pattern expected under a simple synthetic-bias explanation. If OEA were benefiting primarily from GPT-like phrasing, it should dominate all UIQ types. Instead, M2D-CLAP remains strongest on Question, Keyphrase, and Paraphrase averages (Table 4), while OEA's gains are concentrated in Imperative and exclusionary queries. We interpret this as evidence that OEA's strengths arise from instruction-following and semantic discrimination capabilities rather than blanket alignment to generated text style.

J Backbone Generality and Retrieval-Specific Scaling

Reviewer feedback asked whether our findings depend on a single multimodal LLM family. We therefore emphasize that OEA was evaluated with

Model	T2A			T2T		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP Models						
LAION-CLAP	14.10	36.35	48.94	36.41	56.39	65.81
Robust-CLAP	15.94	39.69	52.32	36.40	56.39	65.82
MGA-CLAP	21.06	45.80	57.72	43.52	62.32	70.63
M2D-CLAP	**22.10**	**47.86**	**59.15**	40.43	59.27	68.08
Vanilla LALMs						
Nemotron-3B	5.00	14.64	21.74	32.62	47.47	55.32
Qwen2.5-Omni-3B	0.22	0.82	1.52	21.39	31.79	37.95
Qwen2.5-Omni-7B	0.22	0.90	2.01	22.49	33.97	40.27
OEA Models						
OEA-Nemo3B	21.96	45.91	57.98	44.62	63.36	72.03
OEA-Nemo3B (+Cl)	21.43	46.26	58.38	45.17	64.05	72.47
OEA-Qwen3B	20.88	43.72	55.17	44.94	63.80	72.20
OEA-Qwen3B (+Cl)	21.54	45.43	56.67	46.27	64.47	72.98
OEA-Qwen7B	21.85	46.77	58.56	**46.46**	64.60	73.20
OEA-Qwen7B (+Cl)	21.27	46.40	58.89	46.32	**64.75**	**73.21**

Table 11: Mean baseline results across AudioCaps, Clotho, and MECAT (%). M2D-CLAP achieves best T2A performance, while OEA models substantially outperform all baselines on T2T. **Bold**: best per column, underline: second best.

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
LAION-CLAP	25.13	56.10	71.18	16.17	39.33	53.30	6.84	20.52	31.25
Robust-CLAP	26.15	63.28	77.13	18.09	44.02	57.03	7.43	22.41	31.96
MGA-CLAP	31.90	65.03	78.87	18.18	44.98	57.70	8.25	23.23	36.20
M2D-CLAP	**37.13**	**73.74**	**86.56**	20.00	46.60	58.66	**9.43**	**25.94**	**37.62**
OEA-Nemo3B	36.72	72.00	83.69	19.81	43.54	57.61	7.31	24.41	35.14
OEA-Nemo3B (+Cl)	34.56	65.64	80.31	23.44	50.62	63.73	6.96	22.05	32.43
OEA-Qwen3B	36.72	70.36	82.56	21.34	44.11	59.71	7.43	18.04	27.12
OEA-Qwen3B (+Cl)	35.18	68.31	81.74	**25.74**	**54.26**	**66.79**	6.60	17.57	24.88
OEA-Qwen7B	36.92	70.26	84.72	22.30	48.33	63.06	8.37	23.82	34.79
OEA-Qwen7B (+Cl)	33.23	66.56	79.08	24.11	52.15	66.32	8.02	23.23	34.79

Table 12: Question query T2A results (%). **Bold**: best per column, underline: second best.

three backbones drawn from two organizations: Nemotron-3B (NVIDIA), Qwen2.5-Omni-3B, and Qwen2.5-Omni-7B (Alibaba). Across all three, OEA consistently improves T2T retrieval and remains competitive on T2A, especially on the contamination-free MECAT benchmark.

The scaling trend is also informative. Moving from Qwen3B to Qwen7B yields only marginal or inconsistent changes, and on MECAT T2A R@1 the 7B model slightly underperforms Nemo3B (7.02 vs. 7.96 for the +Cl variants in Table 2). This suggests that audio-text retrieval saturates earlier than generative tasks: once the encoder has sufficient language capacity, further gains depend more on contrastive alignment quality, data curation, and backbone-dataset compatibility than on parameter count alone. In other words, our results point to retrieval-specific scaling behavior rather than a monotonic "larger is always better" law.

K Hard Negative Mining Pipeline

We develop a four-stage pipeline to construct hard negative pairs for exclusionary query evaluation:

Stage 1: Acoustic Candidate Retrieval For each target audio, we retrieve top-K (K=20) acoustically similar candidates using MGA-CLAP audio embeddings.

Stage 2: Acoustic Similarity Filtering We apply a dynamic threshold θ_{acoustic} to retain candidates with high acoustic similarity, keeping approximately $3\times$ the final target count.

Stage 3: Semantic Dissimilarity Scoring We compute semantic similarity between target and

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
LAION-CLAP	22.36	55.38	70.05	14.45	37.99	51.29	5.66	17.81	27.24
Robust-CLAP	26.77	62.36	77.54	16.08	42.58	56.94	5.66	18.28	28.77
MGA-CLAP	34.67	67.59	79.90	17.99	44.59	55.79	8.25	22.64	33.14
M2D-CLAP	33.85	67.49	80.10	20.29	46.22	60.00	7.08	20.52	31.25
OEA-Nemo3B	41.33	**75.08**	**86.36**	18.85	43.25	57.03	8.37	**24.41**	**36.44**
OEA-Nemo3B (+Cl)	37.64	72.51	83.59	24.02	51.29	64.69	7.67	24.41	34.55
OEA-Qwen3B	40.00	72.72	84.92	22.30	46.12	59.33	7.43	19.34	27.00
OEA-Qwen3B (+Cl)	39.49	72.62	84.31	25.45	**55.69**	**67.56**	7.19	17.33	25.12
OEA-Qwen7B	**42.15**	74.77	84.82	24.02	48.71	61.82	**8.49**	24.41	34.55
OEA-Qwen7B (+Cl)	36.72	70.56	83.18	**26.22**	54.64	67.27	8.02	24.41	35.61

Table 13: Imperative query T2A results (%). **Bold**: best per column, underline: second best.

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
LAION-CLAP	28.21	62.97	77.95	15.98	39.14	52.73	6.49	20.64	30.78
Robust-CLAP	30.97	67.90	83.28	18.28	42.58	56.94	6.25	21.46	31.37
MGA-CLAP	38.87	73.54	87.28	19.62	47.56	59.71	**9.43**	25.12	**35.97**
M2D-CLAP	42.36	76.62	**88.92**	21.63	50.72	60.67	7.43	24.41	34.55
OEA-Nemo3B	41.85	**77.85**	88.00	20.96	44.50	57.22	7.43	**25.24**	35.85
OEA-Nemo3B (+Cl)	38.05	73.33	85.13	23.83	50.14	64.50	7.67	23.00	33.96
OEA-Qwen3B	40.10	74.56	86.05	21.24	46.79	60.19	6.49	19.46	26.89
OEA-Qwen3B (+Cl)	39.18	73.03	84.92	**27.56**	**55.50**	**70.81**	5.90	18.63	25.59
OEA-Qwen7B	**42.67**	75.90	87.69	22.78	50.53	61.82	9.08	23.94	35.02
OEA-Qwen7B (+Cl)	38.56	72.10	84.10	26.41	54.83	67.37	7.31	23.82	35.97

Table 14: Paraphrase query T2A results (%). **Bold**: best per column, underline: second best.

candidate captions using BGE-large-en-v1.5 (Xiao et al., 2023) sentence embeddings.

Stage 4: Semantic Dissimilarity Filtering We retain pairs where captions are sufficiently different ($\sim$1$\times$ target count), yielding audio clips that are *acoustically similar but semantically distinct*—ideal hard negatives for exclusionary queries.

Human Verification Because the initial candidate pool is seeded by MGA-CLAP, we do not rely on that model alone to define difficulty. Before finalizing the benchmark, human reviewers inspected sampled target–hard-negative pairs and removed pairs that were not genuinely confusable from an auditory perception standpoint. This additional verification step reduces the risk that the benchmark merely rewards agreement with a single acoustic embedding model.

L Additional Evaluation Metrics

In addition to HNSR@k (reported in main text), we introduce the following metrics for exclusionary query evaluation:

- **Δ-Rank**: $\text{Rank(HN)} - \text{Rank(T)}$. Higher values indicate better separation between target (T) and hard negative (HN). This provides a continuous measure of separation quality suitable for model comparison.
- **HNSR** (Hard Negative Suppression Rate): Percentage of queries where $\text{Rank(HN)} > \text{Rank(T)}$—i.e., target ranked above hard negative, regardless of absolute rank.
- **TFR** (Target-First Rate): Percentage where target is ranked first.
- **TFR-HN@k**: Percentage where target is ranked first AND hard negative is outside top-k—the strictest combined metric for applications requiring high-confidence exclusion understanding.

M Negative Query Detailed Results

Table 17 presents comprehensive negative query evaluation results including all discrimination metrics.

Model	AudioCaps			Clotho			MECAT		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
LAION-CLAP	28.51	62.56	76.92	17.51	41.63	55.02	7.31	23.94	33.49
Robust-CLAP	31.38	67.90	83.28	18.37	47.37	61.44	8.37	26.30	36.44
MGA-CLAP	41.23	76.31	87.08	22.11	50.43	63.64	8.37	26.18	37.85
M2D-CLAP	**46.46**	**82.77**	**93.64**	22.11	48.52	62.49	8.84	**28.18**	**40.21**
OEA-Nemo3B	40.51	73.44	85.44	21.53	45.45	59.14	9.08	27.71	39.74
OEA-Nemo3B (+Cl)	36.92	69.64	83.28	25.93	52.54	66.03	7.90	24.88	37.03
OEA-Qwen3B	38.97	73.03	85.85	24.50	49.76	61.72	7.55	21.82	29.60
OEA-Qwen3B (+Cl)	39.08	72.51	84.51	27.66	**57.99**	**71.87**	6.25	19.69	27.24
OEA-Qwen7B	43.79	74.87	88.00	24.98	50.72	65.93	**9.79**	25.35	36.56
OEA-Qwen7B (+Cl)	38.15	70.46	83.08	**27.75**	55.41	68.61	8.25	24.76	36.20

Table 15: Keyphrase query T2A results (%). **Bold**: best per column, underline: second best.

Model	Audio Encoding ms/clip	Text Encoding ms/query	Peak GPU Memory	Trainable Params
LAION-CLAP	107.7	0.53	∼0.6 GB	158 M
MGA-CLAP	31.7	0.72	∼0.6 GB	148 M
M2D-CLAP	58.1	0.30	0.7 GB	89 M
OEA-Nemo3B	163.8	2.30	11.5 GB	13.7 M
OEA-Qwen3B	539.3	2.60	11.6 GB	16.2 M
OEA-Qwen7B	666.8	4.86	18.3 GB	17.2 M

Table 16: System efficiency benchmark on Clotho. Audio encoding is performed offline and amortized across many user queries, so online serving primarily depends on text encoding latency.

Model	Standard Retrieval		Hard Negative Discrimination			Strict Metrics	
	R@5	R@10	Δ-Rank	HNSR	HNSR@10	TFR	TFR@10
LAION-CLAP	30.72	42.15	25.6	68.7	24.7	10.5	4.9
Robust-CLAP	34.86	46.53	**33.7**	69.6	30.3	11.6	6.9
MGA-CLAP	36.75	48.92	15.1	65.0	21.2	12.9	3.9
M2D-CLAP	**41.56**	**55.63**	18.6	68.0	30.3	16.2	7.5
OEA-Nemo3B	39.64	51.23	21.1	68.6	32.9	15.1	8.6
OEA-Nemo3B (+Cl)	41.48	53.67	17.9	69.2	32.5	**17.3**	**10.1**
OEA-Qwen3B	37.49	49.12	18.8	67.5	31.6	14.3	8.3
OEA-Qwen3B (+Cl)	41.52	53.89	17.6	67.3	32.8	16.4	10.0
OEA-Qwen7B	38.20	50.45	33.0	**74.4**	31.8	15.8	9.5
OEA-Qwen7B (+Cl)	38.14	50.38	29.8	73.9	**34.6**	15.9	9.8

Table 17: Negative query evaluation (mean across datasets). Standard Retrieval: R@k (%). Hard Negative Discrimination: Δ-Rank (rank gap), HNSR/HNSR@10 (%). Strict Metrics: TFR/TFR@10 (%). M2D-CLAP achieves best standard retrieval, while OEA models substantially outperform all baselines on discrimination metrics. **Bold**: best, underline: second best.