MARS-SEP: MULTIMODAL-ALIGNED REINFORCED SOUND SEPARATION

Zihan Zhang[1] **Xize Cheng**[1] **Zhennan Jiang**[2] **Dongjie Fu**[1] **Jingyuan Chen**[1]
Zhou Zhao[1] **Tao Jin**[1*]
[1]**Zhejiang University** [2]**Institute of Automation, Chinese Academy of Sciences**
{zihanzhang.ai,jint_zju}@zju.edu.cn

ABSTRACT

Universal sound separation faces a fundamental misalignment: models optimized for low-level signal metrics often produce semantically contaminated outputs, failing to suppress perceptually salient interference from acoustically similar sources. To bridge this gap, we introduce MARS-Sep, a reinforcement learning framework that reformulates separation as decision making. Instead of simply regressing ground-truth masks, MARS-Sep learns a factorized Beta mask policy that is optimized by a clipped trust-region surrogate with entropy regularization and group-relative advantage normalization. Concretely, we sample masks from a frozen old policy, reconstruct waveforms, and update the current policy using clipped importance ratios—yielding substantially more stable and sample-efficient learning. Multimodal rewards, derived from an audio-text-vision encoder, directly incentivize semantic consistency with query prompts. We further propose a progressive alignment scheme to fine-tune this encoder, boosting its cross-modal discriminability and improving reward faithfulness. Extensive experiments on multiple benchmarks demonstrate consistent gains in Text-, Audio-, and Image-Queried separation, with notable improvements in signal metrics and semantic quality. Our code is available at https://anonymous.4open.science/r/MARS-Sep. Sound separation samples are available at https://mars-sep.github.io/.

1 INTRODUCTION

Sound separation (Liu et al., 2022; Dong et al., 2023; Cheng et al., 2025b; Chen et al., 2022; Mahmud & Marculescu, 2024; Ma et al., 2024; Huang et al., 2025a) is a foundational problem in audio processing with broad impact on downstream tasks such as speech recognition (Shi et al., 2022; Kalda et al., 2024), sound event detection (Turpault et al., 2019; Yin et al., 2025), and acoustic scene analysis (Kim & Chang, 2024; Su et al., 2023). Beyond its standalone importance, separation also serves as a powerful data engine: by decomposing mixtures into isolated sources, it enables large-scale augmentation that improves robustness and generalization (Chiu et al., 2021; Yuan et al., 2022; Cheng et al., 2025a; Manilow et al., 2019). In this work we focus on universal, query-conditioned separation—where the query can be audio, text, or image—and ask how to make the separated output not only signal-clean, but also semantically faithful to the user's intent.

Despite notable progress, prevailing methods are predominantly optimized for distortion or interference-based metrics including SDR, SIR, SAR (Vincent et al., 2006), SI-SDR (Roux et al., 2019) and give limited consideration to semantic alignment during training. This creates a *metric dilemma*: models optimized for waveform reconstruction can score high on signal-level metrics while leaving perceptually salient interference, thus violating semantic correspondence to the query.

To bridge this gap, we propose **MARS-Sep**, a reinforcement learning framework that reformulates mask prediction as stochastic decision-making optimized with multimodal rewards. We cast mask generation as an actor-only trust-region optimization over a factorized Beta policy on time-frequency bins. Training uses a clipped surrogate with entropy regularization and normalized advantages, ensuring stable updates. Instead of focusing on low-level sampling details, our approach leverages multimodal rewards that holistically capture signal fidelity, interference suppression, and perceptual

*Corresponding author.

quality across audio, text, and visual queries. To provide reliable reward signals and mitigate reward hacking, we further introduce a progressive alignment strategy that fine-tunes the multimodal encoder to enhance cross-modal discrimination and stabilize policy learning.

We validate **MARS-Sep** on VGGSOUND-clean+ and MUSIC-clean+ (Dong et al., 2023) across Text-, Audio-, and Image-Queried separation. Extensive experiments show consistent gains over prior methods, improving SDR/SIR/SAR/SI-SDRi and notably higher CLAP score, while qualitative analyses highlight clearer suppression of non-target sources and better category discrimination.

Our contributions are summarized as follows:

- We formulate query-conditioned sound separation as a trust-region reinforcement learning problem, which optimizes a factorized Beta mask policy on time-frequency bins.
- We introduce a progressive alignment strategy that fine-tunes the multimodal encoder to enhance cross-modal discriminability and provide stable, informative reward signals, thereby mitigating reward hacking.
- We demonstrate consistent improvements across SDR/SIR/SAR and SI-SDRi, alongside higher CLAP scores and clearer qualitative separation, confirming both signal-level and semantic gains.

2 Related Work

2.1 Universal and Query-Conditioned Sound Separation

Research on isolating sources from complex mixtures has progressed from domain-specific settings—speech separation (Luo & Mesgarani, 2018; 2019; Zeghidour & Grangier, 2021; Subakan et al., 2021) and music source separation (Luo et al., 2017; Rouard et al., 2023a; Luo & Yu, 2023)—toward Universal Sound Separation (USS) (Kavalerov et al., 2019; Wisdom et al., 2020), which aims to decompose arbitrary mixtures without class constraints. Key enablers include permutation invariant training (PIT) for resolving label permutations (Yu et al., 2017; Postolache et al., 2023) and mixture invariant training (MixIT) for leveraging unlabeled mixtures (Wisdom et al., 2020); large-scale resources such as AudioSet (Gemmeke et al., 2017) further catalyzed progress. Beyond audio-only models, integrating visual context (Majumder et al., 2021; Tan et al., 2023) or using class labels as queries (Chen et al., 2022; 2023) expands separation capability. In parallel, Query-Based Sound Extraction (QBSE) reframes the task as extracting user-specified content while suppressing irrelevant sources. By modality, label-queried systems are simple yet closed-set (Chen et al., 2022; Liu et al., 2024); text-queried approaches such as LASS (Liu et al., 2022) enable open vocabulary but face joint-optimization and generalization hurdles; visual queries exploit images for grounding (Michelsanti et al., 2021; Gao & Grauman, 2021; Ye et al., 2024; Pian et al., 2024); and audio queries use exemplars to target abstract or indescribable sounds (Lee et al., 2019; Chen et al., 2022). Recent attempts unify modalities via cross-attention (Chen et al., 2023) or hybrid encoders (Rouard et al., 2023b), though joint training can limit generalization. Representative systems include CLIPSEP (Dong et al., 2022), which leverages visual data to improve text-queried training, and AudioSep (Liu et al., 2024), which couples CLAP (Wu et al., 2023) with a 14k-hour corpus to achieve strong zero-shot performance. Despite these advances, open-vocabulary robustness, multipolarity operation (extraction and removal within one framework), and scalable multimodal composition (e.g., "dog barking in this image") remain challenging; query-mixup style training (Cheng et al., 2025b) offers a promising direction but still leaves room for improved semantic alignment and stability.

2.2 Reinforcement Learning for Large Language Models

Reinforcement learning from human feedback (RLHF) has become a central paradigm for aligning large language models with human preferences. Early work (Ziegler et al., 2019; Ouyang et al., 2022) established the pipeline of learning a reward model from pairwise preference data and then applying policy optimization with Proximal Policy Optimization (Schulman et al., 2017), which significantly improved controllability and reliability of LLM outputs. Subsequent methods sought to reduce the complexity and instability of this process. Direct Preference Optimization (DPO) (Rafailov et al., 2023), for instance, eliminates the explicit reward model and optimizes a preference-based

objective directly, providing comparable or superior alignment quality with simpler training. More recently, DeepSeek-R1 (Guo et al., 2025) and related studies (Zheng et al., 2025; Yu et al., 2025) explored alternative optimization strategies such as Group Relative Policy Optimization (GRPO), showing improved stability and reasoning capability compared to PPO-based RLHF.

Extensions of these ideas to multimodal systems have further demonstrated the versatility of reinforcement learning for preference alignment. Vision-R1 (Huang et al., 2025b) and R1-VL (Zhang et al., 2025a) applied reinforcement learning with structured rewards to enhance factual grounding and chain-of-thought reasoning in vision-language models. R1-reward (Zhang et al., 2025b) introduced a refined framework for reward modeling that improves semantic fidelity and mitigates reward hacking, thereby strengthening preference alignment across modalities. Complementary efforts, such as Factually Augmented RLHF (Sun et al., 2024) and RLHF-V (Yu et al., 2024), leveraged fine-grained multimodal feedback to reduce hallucinations and improve semantic consistency, highlighting the broader applicability of RLHF beyond text-only domains.

3 METHOD

3.1 PRELIMINARIES

3.1.1 UNIVERSAL SOUND SEPARATION

Universal Sound Separation (USS) is the task of isolating individual sound sources from an arbitrary audio mixture, without prior knowledge of the number or types of sources. Unlike domain-specific separation exemplified by speech enhancement or music source separation, USS aims to generalize across diverse acoustic conditions and sound categories.

Formally, let the observed mixture signal be denoted as $x(t) = \sum_{i=1}^{N} s_i(t)$, where $s_i(t)$ represents the waveform of the i-th underlying source and N is the (unknown) number of sources. The goal of USS is to estimate a set of signals $\{\hat{s}_i(t)\}_{i=1}^{\hat{N}}$ such that each $\hat{s}_i(t)$ corresponds to one of the true sources $s_i(t)$, up to permutation and possibly scaling. That is, $x(t) \approx \sum_{i=1}^{\hat{N}} \hat{s}_i(t)$.

To achieve this, separation models typically operate in the time-frequency domain or directly on the waveform. Let $\mathbf{X} \in \mathbb{C}^{F \times T}$ denote the short-time Fourier transform (STFT) of the mixture, where F and T are the frequency and time dimensions. USS methods aim to construct masks $\{M_i \in [0,1]^{F \times T}\}$ such that $\hat{\mathbf{S}}_i = M_i \odot \mathbf{X}$ with $\odot$ denoting element-wise multiplication. The inverse STFT is then applied to obtain time-domain estimates $\hat{s}_i(t)$.

3.1.2 OMNISEP: UNIFIED OMNI-MODALITY SOUND SEPARATION WITH QUERY-MIXUP

OmniSep (Cheng et al., 2025b) provides the base separation architecture: a frozen ImageBind (Girdhar et al., 2023) encoder maps audio/image/text inputs to a shared feature space and a Separate-Net (U-Net over STFT magnitudes) predicts masks. Let Q_A be the audio query, Q_V be the visual query, and Q_T be the text query. Training in OmniSep mixes queries across modalities via Query-Mixup

$$Q = \frac{w_a Q_A + w_v Q_V + w_t Q_T}{w_a + w_v + w_t}, \quad w_a, w_v, w_t \in [0, 1] \tag{1}$$

and combines intermediate masks into final masks $\hat{M}$ through channel-wise weighting; the supervised objective is the sum of weighted binary cross-entropy (WBCE) losses on ideal masks. OmniSep also supports negative queries by introducing a negative query weight α to adjust the query as $Q' = (1 + \alpha)Q - \alpha Q_N$ to remove interfering content, and, more broadly, frames omni-modal querying with a frozen ImageBind backbone within a unified audio/text/video-queried separation paradigm.

3.2 MARS-SEP: REINFORCEMENT LEARNING FOR MULTI-SOURCE UNIVERSAL SOUND SEPARATION ENHANCEMENT

To advance beyond the deterministic optimization paradigm of OmniSep, we introduce a reinforcement learning framework that reformulates sound separation as a stochastic decision-making problem with multimodal rewards (Figure 1).

Specifically, the separator produces a mask prediction that parameterizes the *new policy* $(\alpha_{new}, \beta_{new})$, while a snapshot from the previous step provides the *old policy* $(\alpha_{old}, \beta_{old})$. Masks are sampled from the old policy to ensure stable training, and the separated audio is compared against audio/text/video queries in a shared embedding space using a fine-tuned ImageBind encoder with a multimodal fusion module. The cosine similarity between the separated audio embedding and the fused query embedding yields a scalar reward, which is normalized by a running baseline and group-relative scaling to form advantages. These are combined with the log-probability ratio between old and new policies (i.e., $\log\pi_{\text{old}}$ and $\log\pi_{\text{new}}$) under a clipped surrogate objective, supplemented by entropy regularization for exploration and a KL penalty for stability. The current policy is then snapshotted to serve as the old policy in the next iteration.

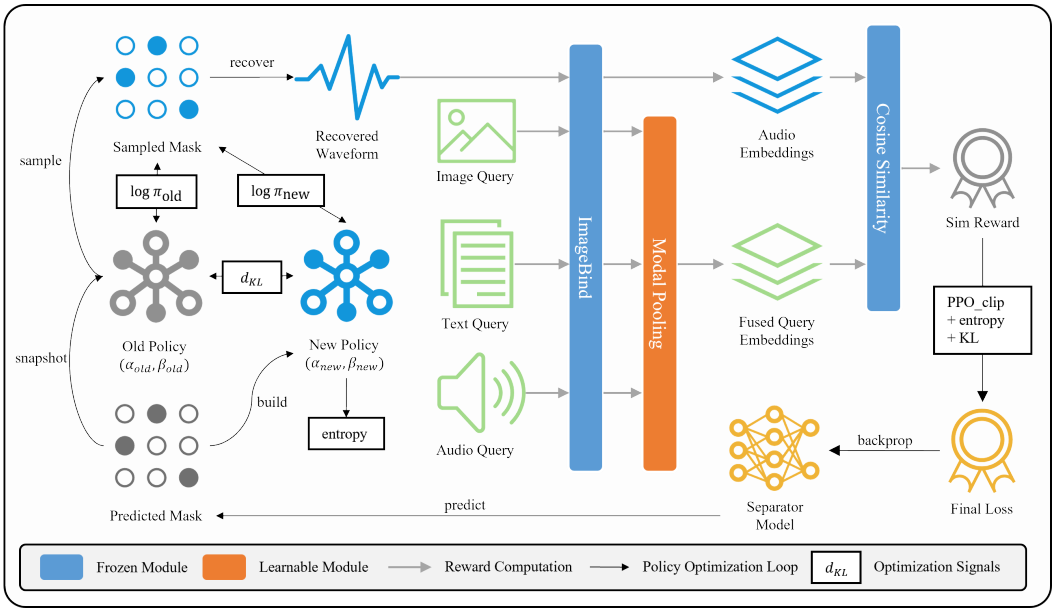

Figure 1: The reinforcement learning loop of MARS-Sep. The separator generates stochastic mask actions from a Beta-distributed policy, while a frozen snapshot serves as the old policy for stable optimization. Multimodal rewards derived from audio, text, and visual embeddings guide policy updates, with entropy and KL regularization enhancing exploration and stability.

3.2.1 Trust-Region-Style Policy Optimization for Stable Mask Sampling

We formulate query-conditioned sound separation as a standard Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$. The state space $\mathcal{S}$ consists of the mixture spectrogram X and the query Q, while the action space $\mathcal{A}$ corresponds to masks M. The transition T is deterministic, $\hat{y} = s(X, M)$, reconstructing the waveform. The reward function R is defined by the similarity between the separated waveform and the multimodal query. Our goal is to train a policy $\pi_\theta(M \mid X, Q)$ that maximizes the cumulative designed reward. Let X be the magnitude spectrogram of a mixture and θ the separator parameters that produce a deterministic mask proposal $P_\theta(X, Q) \in [0, 1]^{H \times W \times K}$, where H is the number of frequency bins, W is the number of time frames and K denotes the number of target sources to be separated. We turn this proposal into a stochastic policy over masks by a factorized Beta distribution

$$\pi_\theta(M \mid X, Q) = \prod_{h,w,k} \text{Beta}\big(M_{h,w,k};\ \alpha_{h,w,k},\ \beta_{h,w,k}\big), \quad \alpha = 1 + \kappa P_\theta,\ \beta = 1 + \kappa(1 - P_\theta), \quad (2)$$

with concentration scale $\kappa > 0$. Reparameterized sampling $M \sim \pi_\theta(\cdot \mid X, Q)$ yields a mask that is close to the proposal yet retains exploration. The masked magnitude is combined with the

mixture phase to reconstruct a waveform $\hat{y} = s(X, M)$; the ground-truth component $y^\star$ is reconstructed analogously from ideal masks (only for reward computation during training). The factorized Beta parametrization aligns with the $[0, 1]$ support of masks and offers a transparent exploration-exploitation knob via the concentration scale κ, which we anneal to avoid degenerate near-binary masks early in training.

At each training step we sample from a frozen old policy $\pi_{\theta_{\text{old}}}$ (a snapshot from the previous step), reconstruct the waveform $\hat{y}$ from the sampled mask M, and compute a scalar reward R against the query-conditioned targets. A moving-average baseline b yields the advantage $A = R - b$. To stabilize updates, we adopt a clipped trust-region style surrogate in the spirit of PPO, using GRPO of the advantage. Concretely, define the importance ratio

$$r_\theta(M) = \frac{\pi_\theta(M \mid X, Q)}{\pi_{\theta_{\text{old}}}(M \mid X, Q)} = \exp\big(\log \pi_\theta(M) - \log \pi_{\theta_{\text{old}}}(M)\big), \tag{3}$$

and let $\tilde{A} = \frac{A - \mu(A)}{\sigma(A) + \varepsilon}$ be the group-relative advantage. The clipped surrogate objective with entropy regularization and KL penalty is

$$\mathcal{J}_{\text{clip}}(\theta) = \mathbb{E}_{M \sim \pi_{\theta_{\text{old}}}}\Big[\min\big(r_\theta(M)\,\tilde{A},\ \text{clip}(r_\theta(M), 1-\epsilon, 1+\epsilon)\,\tilde{A}\big) + \lambda_H\,\mathcal{H}\big(\pi_\theta\big) - \lambda_{\text{KL}}\,\text{KL}\big(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}}\big)\Big], \tag{4}$$

and the loss minimized in training is $\mathcal{L}_{\text{RL}}(\theta) = -\mathcal{J}_{\text{clip}}(\theta)$. Here $\mathcal{H}$ denotes the entropy of the factorized Beta policy, ϵ is the clipping range, and $\lambda_H, \lambda_{\text{KL}} > 0$ control exploration and the trust region, respectively. In practice, $\log \pi$ factorizes over bins; we broadcast $\tilde{A}$ to the mask shape and estimate expectations with Monte Carlo samples per iteration. The old policy $\pi_{\theta_{\text{old}}}$ is updated to the current snapshot after each optimization step, yielding a single-epoch PPO update that preserves the original training loop while markedly improving stability.

This design preserves the benefits of reinforcement learning while avoiding the instability of plain policy gradients, leading to more reliable convergence for mask-based separation. Importantly, it achieves this without introducing additional value networks or complex estimators, keeping the optimization efficient while directly tying policy updates to multimodal reward signals.

3.2.2 MULTIMODAL REWARD

To optimize the separation policy, we define a reward that measures how well the separated audio waveform $\hat{y}$ semantically matches the target query across modalities in a unified embedding space provided by ImageBind. We project audio waveforms, text queries, and sampled video frames into a shared space via ImageBind encoders $\phi_a(\cdot)$, $\phi_t(\cdot)$, and $\phi_v(\cdot)$, respectively.

All embeddings are ℓ_2-normalized, and similarity is measured with cosine similarity:

$$\text{sim}(u, v) = \left\langle \frac{u}{\|u\|}, \frac{v}{\|v\|} \right\rangle. \tag{5}$$

Unimodal rewards. Given separated audio $\hat{y}$, ground-truth audio $y^\star$, a text query embedding $t^\star$, and a video frames embedding $v^\star$, we compute:

$$r_{a\to a} = \text{sim}\big(\phi_a(\hat{y}), \phi_a(y^\star)\big), \quad r_{t\to a} = \text{sim}\big(\phi_a(\hat{y}), \phi_t(t^\star)\big), \quad r_{v\to a} = \text{sim}\big(\phi_a(\hat{y}), \phi_v(v^\star)\big). \tag{6}$$

These terms measure acoustic fidelity, semantic alignment with text, and consistency with visual context, respectively.

Aggregation strategy: Query-pooling. Instead of comparing unimodal similarities separately, we fuse the target-side multimodal embeddings into a joint representation using Multi-Modal Low-Rank Bilinear Pooling (MLBP) [1] ((Kim et al., 2017)) and compare it directly to the separated audio

[1] See Appendix A for a detailed description.

embedding. Specifically, $z^\star = \mathrm{MLBP}\big(\phi_a(y^\star), \phi_t(t^\star), \phi_v(v^\star)\big)$, and the scalar reward is $R = \mathrm{sim}\big(\phi_a(\hat{y}), z^\star\big)$.

The motivation for pooling is to ensure the reward captures joint multimodal consistency rather than independent unimodal matches. Audio, text, and vision carry complementary cues: audio encodes acoustic details, text conveys semantic categories, and vision provides environmental context. If each is compared separately, the reward may overweight a single modality. By applying low-rank bilinear pooling, we explicitly model multiplicative interactions between modalities (e.g., a textual query specifying an instrument that also appears visually). This fused target anchor $z^\star$ encourages the separated audio to simultaneously align with all modalities, yielding more semantically faithful and robust rewards. This asymmetric design mirrors the implementation: the separated audio remains in its native representation while the target modalities are fused into a semantic anchor. This reduces variance from stochastic mask sampling and provides a stable training signal.

In the next section, we describe the progressive fine-tuning curriculum used to initialize this policy with robust cross-modal alignment before RL optimization.

3.3 Multimodal Encoder Fine-tuning via Progressive Alignment

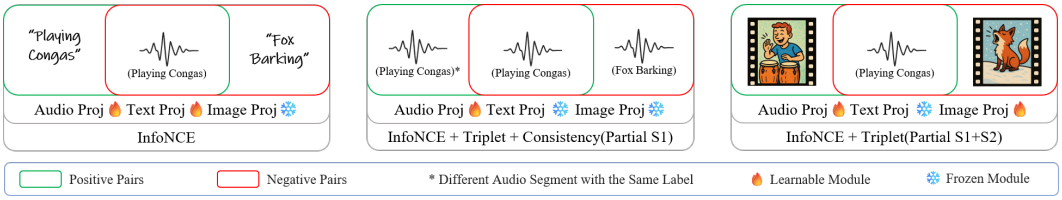

Figure 2: Progressive fine-tuning strategy for sound source discrimination and separation. Encoders remain frozen while task-specific heads are gradually unfrozen and each stage builds on the best checkpoint from the previous one. The two latter stages are trained with a fraction of the former aligned paired data to avoid catastrophic forgetting.

To enhance discrimination between same- and different-source signals, we apply a multimodal contrastive fine-tuning objective on ImageBind: for each audio segment, positives pair it with an audio clip, video frame, or label text of the same class, while negatives pair it with content from other classes; by optimizing a contrastive loss over these multimodal pairs, the model is encouraged to bring embeddings of semantically consistent sources closer together, while pushing apart those belonging to different sound categories.

The fine-tuning process, as shown in Figure 2, is organized into three sequential stages, each with a distinct training objective, and each stage begins from the best-performing checkpoint obtained in the previous one. This curriculum design allows the model to move gradually from semantic grounding to intra-class discrimination and finally to multimodal alignment.

3.3.1 Sample Pair Construction for Contrastive Learning

We build multimodal contrastive pairs with the audio clip as the anchor. Positives pair the anchor with (i) its label text, (ii) another audio instance from the same class, or (iii) frames from the temporally aligned video segment. Negatives use labels, audio, or frames from other classes or temporally mismatched segments. This unified scheme pulls matched embeddings together while pushing mismatched ones apart.

In the first stage, the model is trained to align audio signals with their corresponding textual labels. At this point, all modality encoders (audio, text, vision) and postprocessors are kept frozen to preserve their pretrained representations, while only the projection heads together with a shared temperature parameter are unfrozen and updated. The training objective is a symmetric InfoNCE loss, which encourages paired audio-text embeddings to converge while repelling mismatched pairs. Formally, for a batch of size N, with normalized embeddings z_a^i and z_t^i, the loss is defined as

$$\mathcal{L}_{\mathrm{S1}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\langle z_a^i, z_t^i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle z_a^i, z_t^j\rangle/\tau)} + \log\frac{\exp(\langle z_t^i, z_a^i\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle z_t^i, z_a^j\rangle/\tau)}\right], \tag{7}$$

where τ is a learnable temperature scaling factor. This stage establishes the initial semantic grounding of audio in the shared embedding space while limiting parameter updates to lightweight layers.

The second stage focuses on audio-audio discrimination. Again, all encoders and postprocessors remain frozen, while the audio projection head and shared temperature are unfrozen to adapt representations for finer discrimination. Given an anchor audio clip, another clip of the same class is selected as the positive, while a clip of a different class serves as the negative. The objective combines several terms: an InfoNCE loss to enforce alignment between same-class pairs, a triplet loss to guarantee a margin between positive and negative similarities, and a consistency loss to ensure invariance to perturbations. Specifically,

$$\mathcal{L}_{\mathrm{S2}} = \lambda_1\,\mathcal{L}_{\mathrm{InfoNCE}}(z_1, z_2) + \lambda_2\,\max(0, [1-\cos(z_1, z_2)] - [1-\cos(z_1, z_n)] + m) + \lambda_3\|z_1 - z_2\|^2, \tag{8}$$

where z_1 and z_2 are embeddings of audio from the same class, z_n is a negative sample, and m is a margin hyperparameter. To mitigate catastrophic forgetting, a fraction of audio-text pairs from stage one are mixed into training, ensuring that semantic alignment is preserved while discrimination improves.

The third stage introduces visual grounding through audio-video pairs. All encoders and postprocessors remain frozen, but the audio and vision projection heads are unfrozen and trained jointly, along with the shared temperature. Uniformly sampled frames from the corresponding video provide the positive modality, while frames from other videos or temporally misaligned portions serve as negatives. The objective again includes an InfoNCE term between audio and video embeddings, along with a triplet loss that uses mismatched video frames as hard negatives. To maintain previously acquired capabilities, the objectives from stage one and stage two are partially incorporated. The overall loss can be expressed as

$$\mathcal{L}_{\mathrm{S3}} = \mu_1\,\mathcal{L}_{\mathrm{InfoNCE}}(z_a, z_v^+) + \mu_2\,\mathcal{L}_{\mathrm{Triplet}}(z_a, z_v^+, z_v^-) + \mu_3\,\mathcal{L}_{\mathrm{S1}} + \mu_4\,\mathcal{L}_{\mathrm{S2}}, \tag{9}$$

where z_v^+ and z_v^- are embeddings of positive and negative video samples, and the coefficients μ_i balance the relative contributions.

At the end of stage 1 and 2, the best checkpoint is used to initialize the subsequent stage. Stage one therefore provides a semantic anchor via audio-text alignment, stage two sharpens class discrimination through audio-audio comparison, and stage three consolidates multimodal grounding by linking audio with vision. This progressive fine-tuning procedure, in which encoders are kept frozen and only task-relevant heads and scaling parameters are successively unfrozen, ensures that the model evolves in a stable and interpretable manner, acquiring increasingly sophisticated capabilities for sound source discrimination and separation.

4 Experiments

4.1 Experiment Setup

We evaluate our approach on two widely used audio-visual separation benchmarks, **VGGSound** (Chen et al., 2020) and **MUSIC** (Zhao et al., 2018). VGGSound is a large-scale dataset with over 300 sound categories collected from YouTube videos, offering substantial acoustic and visual diversity; MUSIC is a smaller dataset of solo and duet music performance videos spanning a variety of instruments, which emphasizes structured harmonic signals and thus provides a complementary and cross-domain setting. Training and data preprocessing details are provided in Appendix B.

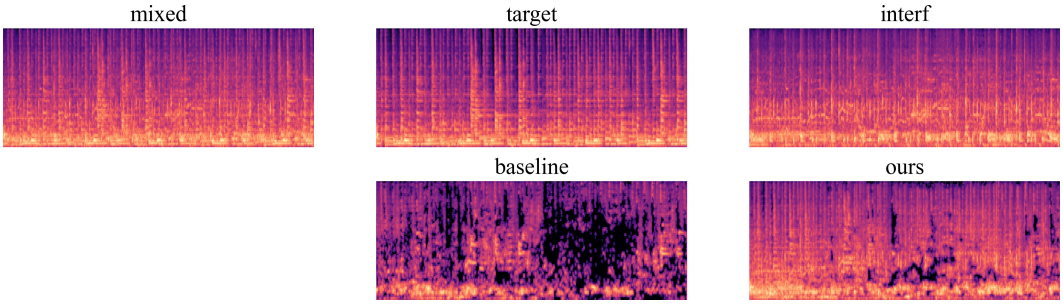

Figure 3: Log-mel spectrograms of separated audio from different query modalities on VGGSOUND-clean+ dataset. The target source is "cattle bovinae cowbell". From left to right: (a) Mixture of "cattle bovinae cowbell" and "tap dancing"; (b) Ground-truth "cattle bovinae cowbell"; (c) Interference "tap dancing"; (d) Separation with text query by the baseline model; (e) Separation with text query by our model.

We adopt standard separation metrics for evaluation, including Signal-to-Interference Ratio (SIR), Signal-to-Distortion Ratio (SDR), Signal-to-Artifact Ratio (SAR), and Scale-Invariant SDR improvement (SI-SDRi)[2].

We additionally report the **CLAP score**, which measures the semantic consistency between the separated audio and its textual label using a contrastive language-audio pretraining model. While traditional signal-level metrics evaluate separation quality in terms of distortion, interference suppression, and artifact reduction, the CLAP score complements them by capturing whether the separated waveform preserves the intended semantic content. Together, these metrics provide a comprehensive assessment of both perceptual signal fidelity and semantic correctness of the separation output.

We select three representative baselines for comparison. **CLIPSep-NIT** (the noise-invariant training version released by the authors) employs CLIP embeddings to guide separation with either visual or textual cues. **AudioSep** adopts large-scale training with text queries to achieve strong open-domain generalization. **OmniSep** integrates multiple modalities into a unified separation framework, highlighting the potential of multimodal fusion. These baselines jointly cover vision-guided, text-guided, and multimodal approaches, providing a comprehensive and challenging reference for evaluating the effectiveness and generalizability of our method.

4.2 Main Results

We first present results on the **VGGSound-clean+** dataset, a refined subset of VGGSound that filters out noisy annotations and ensures higher-quality alignment between audio and visual streams. *MARS-Sep* demonstrates the strongest overall performance across text, audio, image, and composed queries (Table 1). It attains the highest SDR and CLAP score across modalities and achieves the best SI-SDRi in three settings (tied for best under audio queries). On a subset of measures, the balance shifts toward *OmniSep* with notably higher SIR for audio queries and higher SAR for image and composed queries, though the margins are modest. Taken together, these results indicate that reinforcement learning with multimodal rewards improves semantic fidelity and signal quality on balance, while remaining competitive or superior on the remaining metrics.

To further validate cross-domain generalizability, we evaluate on **MUSIC-clean+**, which is derived from the MUSIC dataset and focuses on solo and duet instrumental performances. Compared with VGGSound, MUSIC emphasizes structured harmonic and timbral patterns rather than broad acoustic diversity, making it a complementary benchmark. As reported in Table 2, *MARS-Sep* again achieves clear gains over competing approaches under all query modalities, especially notably higher CLAP scores than all baselines, highlighting that our separated signals preserve semantic consistency with the intended text source, confirming that our method not only handles open-domain separation but also excels in structured, music-centric scenarios.

[2]A detailed description of the SI-SDR and SI-SDRi calculation procedure is presented in Appendix C.

Table 1: Comparison of sound separation performance among different methods on **VGGSOUND-clean+** dataset. Metrics include SIR, SDR, SAR, and SI-SDRi (all in dB), and CLAP score (%).

Methods	VGGSOUND-clean+				
	Mean SDR↑	Mean SIR↑	Mean SAR↑	Mean SI-SDRi↑	Mean $CLAP_t$ ↑
Text Query Sound Separation					
CLIPSEP-NIT (Dong et al., 2023)	2.71±0.87	4.58±1.37	13.60±0.68	2.41±0.53	7.97±0.94
AudioSep (Liu et al., 2022)	6.26±0.87	8.69±0.90	12.85±0.92	4.01±0.59	8.21±0.96
OmniSep (Cheng et al., 2025b)	6.70±0.66	9.04±0.98	13.61±0.77	4.38±0.48	8.98±0.89
MARS-Sep (ours)	**6.91±0.68**	**9.14±1.00**	**13.73±0.77**	**4.55±0.44**	**9.03±0.94**
Audio Query Sound Separation					
OmniSep (Cheng et al., 2025b)	7.15±0.65	**11.65±1.02**	11.84±0.81	4.35±0.52	8.60±0.91
MARS-Sep (ours)	**7.33±0.67**	11.63±1.00	**12.00±0.84**	**4.36±0.50**	**8.91±0.91**
Image Query Sound Separation					
CLIPSEP-NIT (Dong et al., 2023)	4.61±0.82	8.11±1.32	12.06±0.78	3.48±0.60	8.50±0.92
OmniSep (Cheng et al., 2025b)	6.66±0.65	10.00±1.05	**13.73±0.76**	4.43±0.50	8.79±0.89
MARS-Sep (ours)	**6.93±0.67**	**10.18±1.04**	13.41±0.72	**4.57±0.47**	**9.19±0.91**
Composed Omni-modal Query Sound Separation					
OmniSep (Cheng et al., 2025b)	7.79±0.72	**10.76±1.00**	**14.53±0.93**	5.16±0.47	8.85±0.92
MARS-Sep (ours)	**7.93±0.75**	10.65±1.00	14.49±0.95	**5.20±0.45**	**9.22±0.90**

Table 2: Comparison of sound separation performance among different methods on **MUSIC-clean+** dataset.

Methods	MUSIC-clean+				
	Mean SDR↑	Mean SIR↑	Mean SAR↑	Mean SI-SDRi↑	Mean $CLAP_t$ ↑
Text Query Sound Separation					
CLIPSEP-NIT (Dong et al., 2023)	11.03±0.98	16.40±1.38	17.37±0.97	7.53±0.90	5.29±0.96
AudioSep (Liu et al., 2022)	11.23±0.92	16.90±1.31	17.29±0.90	8.56±0.84	5.48±1.02
OmniSep (Cheng et al., 2025b)	12.37±0.85	17.51±1.16	17.96±0.90	9.18±0.79	5.41±0.98
MARS-Sep (ours)	**12.91±0.93**	**17.61±1.17**	**18.28±0.93**	**9.85±0.82**	**5.18±0.01**
Audio Query Sound Separation					
OmniSep (Cheng et al., 2025b)	10.37±0.86	17.76±1.05	14.51±0.88	7.18±1.07	5.39±1.01
MARS-Sep (ours)	**11.73±0.88**	**19.65±1.14**	**15.25±0.86**	**8.38±1.03**	**5.64±1.06**
Image Query Sound Separation					
CLIPSEP-NIT (Dong et al., 2023)	11.64±0.98	18.40±1.26	17.04±1.05	8.27±0.94	5.97±0.94
OmniSep (Cheng et al., 2025b)	13.03±0.96	18.97±1.16	17.88±1.00	10.21±0.89	6.53±1.03
MARS-Sep (ours)	**13.64±1.06**	**19.24±1.16**	**18.05±1.06**	**10.70±0.89**	**6.94±1.06**
Composed Omni-modal Query Sound Separation					
OmniSep (Cheng et al., 2025b)	13.29±0.96	19.55±1.17	17.88±0.96	10.22±0.89	6.35±1.05
MARS-Sep (ours)	**13.89±0.98**	**19.90±1.18**	**17.99±0.97**	**10.78±0.81**	**6.82±0.99**

We provide log-mel spectrograms of a representative sample from the test set of VGGSOUND in Figure 3 for result visualization. Compared with the baseline, MARS-Sep suppresses non-target components more selectively, preserving the target's harmonic ridges and temporal continuity instead of the blocky dropouts visible in the baseline spectrogram. More samples are shown in Appendix E.4.

5 Conclusion

We present MARS-Sep, a multimodal-aligned reinforced sound separation approach that frames sound separation as stochastic decision-making guided by multimodal rewards, enforcing semantic consistency with audio, text, and visual queries rather than optimizing only signal-level metrics. Built on a trust-region style policy with progressive alignment of multimodal encoders, it achieves stable training and strong cross-modal discrimination. Experiments on VGGSOUND-clean+ and MUSIC-clean+ show consistent improvements in fidelity and semantic alignment, advancing semantically aware separation that better matches perceptual quality.

REFERENCES

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14675–14686, 2023.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Zero-shot audio source separation through query-based learning from weakly-labeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4441–4449, 2022.

Xize Cheng, Slytherin Wang, Zehan Wang, Rongjie Huang, Tao Jin, and Zhou Zhao. Unleashing the power of natural audio featuring multiple sound sources, 2025a. URL `https://arxiv.org/abs/2504.17782`.

Xize Cheng, Siqi Zheng, Zehan Wang, Minghui Fang, Ziang Zhang, Rongjie Huang, Shengpeng Ji, Jialong Zuo, Tao Jin, and Zhou Zhao. Omnisep: Unified omni-modality sound separation with query-mixup. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Ching-Yu Chiu, Joann Ching, Wen-Yi Hsiao, Yu-Hua Chen, Alvin Wen-Yu Su, and Yi-Hsuan Yang. Source separation-based data augmentation for improved joint beat and downbeat tracking. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 391–395, 2021. doi: 10.23919/EUSIPCO54536.2021.9616022.

Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick. Clipsep: Learning text-queried sound separation with noisy unlabeled videos. In *The Eleventh International Conference on Learning Representations*, 2023.

Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15490–15500. IEEE, 2021.

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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, Jing Bi, Wenqiang Liu, Nima Mesgarani, and Chenliang Xu. Zerosep: Separate anything in audio with zero training. *arXiv preprint arXiv:2505.23625*, 2025a.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025b.

Joonas Kalda, Ricard Marxer, Tanel Alumäe, Hervé Bredin, et al. Pixit: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings. In *Proc. odyssey 2024*, pp. 115–122, 2024.

Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R. Hershey. Universal sound separation. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 175–179, 2019. doi: 10.1109/WASPAA.2019.8937253.

Jin Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung Woo Ha, and Byoung Tak Zhang. Hadamard product for low-rank bilinear pooling. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

Yungyeo Kim and Joon-Hyuk Chang. Acoustic-scene-aware target sound separation with sound embedding refinement. *IEEE Access*, 12:71606–71616, 2024. doi: 10.1109/ACCESS.2024.3402736.

Jie Hwan Lee, Hyeong-Seok Choi, and Kyogu Lee. Audio query-based music source separation. *arXiv preprint arXiv:1908.06593*, 2019.

Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Separate what you describe: Language-queried audio source separation. In *Proc. Interspeech 2022*, pp. 1801–1805, 2022.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Yi Luo and Nima Mesgarani. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700, 2018. doi: 10.1109/ICASSP.2018.8462116.

Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8): 1256–1266, 2019. doi: 10.1109/TASLP.2019.2915167.

Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1893–1901, 2023. doi: 10.1109/TASLP.2023.3271145.

Yi Luo, Zhuo Chen, John R Hershey, Jonathan Le Roux, and Nima Mesgarani. Deep clustering and conventional networks for music separation: Stronger together. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 61–65. IEEE, 2017.

Hao Ma, Zhiyuan Peng, Xu Li, Mingjie Shao, Xixin Wu, and Ju Liu. Clapsep: Leveraging contrastive pre-trained model for multi-modal query-conditioned target sound extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Tanvir Mahmud and Diana Marculescu. Opensep: Leveraging large language models with textual inversion for open world audio separation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13244–13260, 2024.

Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 275–285, 2021.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. pp. 45–49, 10 2019. doi: 10.1109/WASPAA.2019.8937170.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1368–1396, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Weiguo Pian, Yiyang Nan, Shijian Deng, Shentong Mo, Yunhui Guo, and Yapeng Tian. Continual audio-visual sound separation. *Advances in Neural Information Processing Systems*, 37:76058–76079, 2024.

Emilian Postolache, Jordi Pons, Santiago Pascual, and Joan Serrà. Adversarial permutation invariant training for universal sound separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096847.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023b. doi: 10.1109/ICASSP49357.2023.10096956.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jing Shi, Xuankai Chang, Shinji Watanabe, and Bo Xu. Train from scratch: Single-stage joint training of speech separation and recognition. *Computer Speech & Language*, 76:101387, 2022.

Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305. Springer, 2018.

Yiyang Su, Ali Vosoughi, Shijian Deng, Yapeng Tian, and Chenliang Xu. Separating invisible sounds toward universal audiovisual scene-aware sound separation. *arXiv preprint arXiv:2310.11713*, 2023.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25. IEEE, 2021.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. Language-guided audio-visual source separation via trimodal consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10575–10584, 2023.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. URL `https://hal.inria.fr/hal-02160855`.

Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J. Weiss, Kevin Wilson, and John R. Hershey. Unsupervised sound separation using mixture invariant training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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

Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, Shuhan Qi, Kejia Zhang, Jianyuan Sun, and Wenwu Wang. A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining. *arXiv preprint arXiv:2407.04936*, 2024.

Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5508–5519, 2024.

Han Yin, Jisheng Bai, Yang Xiao, Hui Wang, Siqi Zheng, Yafeng Chen, Rohan Kumar Das, Chong Deng, and Jianfeng Chen. Exploring text-queried sound event detection with audio source separation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017. doi: 10.1109/ICASSP.2017.7952154.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Siyuan Yuan, Zhepei Wang, Umut Isik, Ritwik Giri, Jean-Marc Valin, Michael M. Goodwin, and Arvindh Krishnaswamy. Improved singing voice separation with chromagram-based pitch-aware remixing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 111–115, 2022. doi: 10.1109/ICASSP43922.2022.9747612.

Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.

Yi-Fan Zhang, Xingyu Lu, Xiao Hu, Chaoyou Fu, Bin Wen, Tianke Zhang, Changyi Liu, Kaiyu Jiang, Kaibing Chen, Kaiyu Tang, et al. R1-reward: Training multimodal reward model through stable reinforcement learning. *arXiv preprint arXiv:2505.02835*, 2025b.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A MULTI-MODAL LOW-RANK BILINEAR POOLING (MLBP).

Based on (Kim et al., 2017), let there be K modalities with input vectors $x^{(k)} \in \mathbb{R}^{d_k}$. Each is projected to a shared dimension d via linear transformations without bias:

$$\tilde{x}^{(k)} = W_k x^{(k)}, \quad W_k \in \mathbb{R}^{d \times d_k}. \tag{10}$$

The projected vectors are then fused by an element-wise Hadamard product:

$$p = \bigodot_{k=1}^{K} \tilde{x}^{(k)} \in \mathbb{R}^d. \tag{11}$$

Finally, an output projection with bias produces the pooled embedding:

$$z = W_o p + b, \quad W_o \in \mathbb{R}^{d \times d},\ b \in \mathbb{R}^d. \tag{12}$$

This design captures higher-order modality interactions in a compact manner.

B EXPERIMENTAL SETUP

Following (Dong et al., 2023; Cheng et al., 2025b), for all audio samples, we conducted experiments on samples of length 65535 (approximately 4 seconds) at a sampling rate of 16 kHz. For spectrum computation, we employed a short-time Fourier transform (STFT) with a filter length of 1024, a hop length of 256, and a window size of 1024. All images were resized to 224×224 pixels. The audio model in this paper is a widely used 7-layer U-Net network with $k = 32$, generating 32 intermediate masks. All models were trained with a batch size of 128, using the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\sigma = 10^{-8}$, for 200,000 steps. Additionally, we employed warm-up and gradient clipping strategies, following (Dong et al., 2023). We compute the signal-to-distortion ratio (SDR) using museval (Stöter et al., 2018). All experiments were conducted on a single A100 GPU with 40GB display memory.

C SI-SDR AND SI-SDRI — IMPLEMENTATION

We evaluate separation quality per utterance in the time domain, reconstructing both mixture and estimates via iSTFT with the same analysis parameters used in training. For each utterance with N sources we form 1-D waveforms for the mixture x, references $\{s_k\}_{k=1}^N$, and estimates $\{\hat{s}_k\}_{k=1}^N$. All signals are cropped to the common minimum length L and zero-meaned prior to scoring; SI-SDR calculations run in "float64" for numerical stability. An energy guard is applied: if the absolute sum of any reference or its matched estimate is $\leq 10^{-5}$, the utterance is excluded from SI-SDR/-i aggregation and we report the number of skipped items. No temporal delay search is performed—i.e., we assume sample-level alignment from the dataset and the iSTFT pipeline.

For a single reference-estimate pair $(s, \hat{s})$ (zero-mean, length L), SI-SDR follows the standard scale-invariant projection:

$$\alpha = \langle \hat{s}, s \rangle / \|s\|_2^2, s_{\text{target}} = \alpha s, e = \hat{s} - s_{\text{target}}, \text{SI-SDR}(\hat{s}, s) = 10 \log_{10} \left(\|s_{\text{target}}\|_2^2 / \|e\|_2^2 \right) \tag{13}$$

No filtering beyond the scalar α is allowed. In the multi-source case we build the matrix $S \in \mathbb{R}^{N \times N}$ with entries $S_{k,n} = \text{SI-SDR}(\hat{s}_n, s_k)$ and obtain the permutation that maximizes the total SI-SDR via the Hungarian algorithm applied to $-S$; the resulting per-source SI-SDR values under this assignment are recorded alongside the corresponding mixture baselines.

Improvement is measured against the unaltered mixture. For each reference s_k we compute $\text{SI-SDR}(x, s_k)$ using the same preprocessing, and define the sample-level SI-SDRi as the mean per-source gain under the optimal assignment $\pi^\star$:

$$\text{SI-SDR}i = \frac{1}{N}\sum_{k=1}^{N}\Big[\text{SI-SDR}(\hat{s}_{\pi^\star(k)}, s_k) - \text{SI-SDR}(x, s_k)\Big]. \tag{14}$$

We also store per-source SI-SDR and mixture-baseline SI-SDR lists for analysis. Dataset-level scores are then obtained by averaging per-utterance values; we additionally report either standard deviation or a 95% bootstrap confidence interval. When breaking down by categories (e.g., query modality), we aggregate within category first and macro-average across categories.

D Reinforcement Learning Training Details

Policy and sampling. The separator outputs mask proposals parameterizing a factorized Beta policy $\pi_\theta(M \mid X, Q) = \prod_{t,f,c} \text{Beta}\big(\alpha_{tfc}, \beta_{tfc}\big)$ with $(\alpha, \beta) = \big(1 + 9m,\ 1 + 9(1-m)\big)$ from the network logits $m \in [0,1]$. At each iteration we sample M from a frozen old policy $\pi_{\theta_{\text{old}}}$ (a one-step snapshot) and reconstruct the waveform via iSTFT using the same analysis parameters as in training.

Objective. We maximize a PPO-style clipped surrogate with entropy regularization and an optional KL penalty:

$$\mathcal{J}_{\text{clip}}(\theta) = \mathbb{E}_{M\sim\pi_{\theta_{\text{old}}}}\Big[\min\big(r_\theta(M)\tilde{A},\ \text{clip}(r_\theta(M), 1-\epsilon, 1+\epsilon)\tilde{A}\big) + \lambda_H\,\mathcal{H}(\pi_\theta) - \lambda_{\text{KL}}\,\text{KL}(\pi_\theta \| \pi_{\theta_{\text{old}}})\Big],$$

where $r_\theta(M) = \exp(\log \pi_\theta - \log \pi_{\theta_{\text{old}}})$ and $\tilde{A}$ is the advantage after baseline subtraction and, when enabled, group-relative normalization. We minimize $\mathcal{L}_{\text{RL}}(\theta) = -\mathcal{J}_{\text{clip}}(\theta)$ and update $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ after each step (single-epoch PPO).

Advantages and baselines. Rewards are the cosine similarities between separated audio embeddings and query-conditioned targets (audio/text/video or their mixup/adaptive variants). We use an EMA baseline $b \leftarrow \beta b + (1-\beta)\,\mathbb{E}[R]$ with $\beta = 0.92$. GRPO normalization (optional) sets $\tilde{A} = (A - \mu(A))/(\sigma(A) + 10^{-6})$ within the current group.

Default hyperparameters. Clipping range $\epsilon = 0.2$; entropy coefficient $\lambda_H = 0.1$; KL coefficient $\lambda_{\text{KL}} \in \{0, 0.01\}$ (on by default); one Monte Carlo sample per step; mixed precision (FP16/BF16) for the separator, FP32 for reward computation; AdamW with learning rate 2×10^{-4}, weight decay 0.01; global batch size B as reported in the main text; gradient clipping at 1.0; early stopping on validation reward. Unless otherwise noted, GRPO is enabled.

Reward encoder alignment. We apply progressive alignment of the multimodal encoder prior to RL (staged contrastive objectives with the encoder trunk largely frozen and projection heads trainable), then keep the encoder frozen during RL unless specified. This improves reward faithfulness and stability.

Evaluation protocol. Permutation-invariant matching is used for multi-source cases; SI-SDR/-i are computed per utterance with mixture as the improvement baseline and no delay search, following our metric appendix. All systems share identical evaluation knobs.

E More Experiments

E.1 Ablation Studies

E.1.1 Effect of Progressive Fine-tuning on Source Discrimination of ImageBind

To verify that **progressive fine-tuning improves ImageBind's ability to discriminate between target and non-target sounds**, for each target audio sample in the VGGSOUND test set (with its corresponding text label), we randomly selected an interference audio with a different label and constructed a mixture of the two. We then compared the cosine similarity between the target text

embedding and the embeddings of both the clean target and the mixture, using the pretrained and fine-tuned ImageBind models. By averaging the similarity differences $\text{sim}(\text{emb}_{target}, \text{emb}_{mixture})$ across all test samples, we obtain a robust measure of the model's ability to discriminate non-target sounds. The results in Table 3 demonstrate that the fine-tuned model consistently yields a larger average difference, confirming its improved semantic alignment in the presence of interfering sources.

Table 3: Average similarity differences (target - mixture) relative to the target text embedding, evaluated across the full test set. Larger values indicate stronger discrimination of non-target sources.

Model	**Avg. Difference (↑)**
Pretrained ImageBind	0.0035 ± 0.0561
Fine-tuned ImageBind	$\mathbf{0.0258 \pm 0.0630}$

E.1.2 Effect of reinforcement learning and encoder fine-tuning under different training pipelines.

To disentangle the contributions of reinforcement learning (RL) and progressive fine-tuning (FT) of the ImageBind encoder, we compared four training configurations: (i) baseline supervised training with a frozen encoder, (ii) RL with frozen encoder, (iii) FT-only under supervised training, and (iv) RL combined with FT (our full model). Results on the test set in Table 4 reveal a consistent trend. The FT-only variant yields higher SAR scores but substantially lower SDR, SIR, SI-SDRi and CLAP score, indicating that the encoder becomes more sensitive to semantic cues but the conventional objective fails to enforce clean separation, leading to leakage from interfering sources. By contrast, the RL-only variant achieves improvements over the baseline across all metrics, demonstrating that policy optimization itself enhances separation fidelity even without encoder adaptation. Finally, the RL+FT variant provides the best overall performance, simultaneously improving SDR/SIR and achieving the highest SAR and CLAP scores. These findings confirm that reinforcement learning is crucial for harnessing the benefits of fine-tuned encoders while avoiding the metric trade-off observed in the FT-only setting.

Table 4: Comparison of different training configurations on the VGGSOUND-clean+ test set with text queries. RL here stands for reinforcement learning and FT denotes progressive fine-tuning of ImageBind.

Method	Mean SDR↑	Mean SIR↑	Mean SAR↑	Mean SI-SDRi↑	Mean CLAP_t ↑
Baseline (Supervised + Frozen Encoder)	6.70±0.66	9.04±0.98	13.61±0.77	4.38±0.48	8.98±0.89
RL-only (RL + Frozen Encoder)	6.71±0.70	9.04±1.02	14.08±0.80	4.50±0.75	8.96±0.90
FT-only (Supervised + Fine-tuned Encoder)	0.75±0.64	1.41±1.18	**87.13±0.15**	0.00±0.00	5.48±0.95
RL+FT (Full Model)	**6.91±0.68**	**9.14±1.00**	13.73±0.77	**4.55±0.44**	**9.03±0.94**

This case further illustrates that while fine-tuning enhances the encoder's semantic sensitivity, reinforcement learning is indispensable to suppress residual noise and achieve clean separation. In combination, RL and FT strike a balance between semantic alignment and signal fidelity, yielding perceptually superior outputs.

E.2 Resolving Semantic Ambiguity in Acoustically Similar Sources

To qualitatively evaluate our model's ability to address the "metric dilemma", we designed a case study focused on separating acoustically similar sources. We created a challenging audio mixture containing both the sound of **tap dancing** and **typewriting** simultaneously. These sources are highly confusable as both are characterized by sharp, percussive transients with broadband spectral content, lacking the strong, sustained harmonic structures that typically aid in separation.

The task was to isolate the tap dancing using the text query "the sound of tap dancing." When this mixture was processed by the baseline OmniSep model (without RL), it achieved a high Signal-to-Distortion Ratio (SDR), yet the resulting audio was perceptually contaminated with the distinct, rhythmic clicks of the typewriter. This outcome exemplifies the metric dilemma, where a model successfully optimizes a signal-level metric while failing to achieve true semantic separation.

In contrast, our proposed MARS-Sep, guided by a multimodal semantic reward, produced a much cleaner separation. While its SDR score was marginally lower, its SIR was substantially higher, indicating superior suppression of the interfering typewriter source. To further quantify this semantic improvement, we computed the CLAP score ((Xiao et al., 2024)), defined as **the cosine similarity between the separated audio's embedding and the text query's embedding** using the CLAP model. Unlike purely signal-level metrics, the CLAP score directly measures semantic alignment across modalities, offering a more reliable indicator of whether the separated source matches the intended textual description. The comparative results are summarized in Table 5.

Table 5: A quantitative and qualitative comparison for the "tap dancing"-"typewriting" separation task. This table presents the results for the baseline OmniSep model and our proposed MARS-Sep. The CLAP score is the cosine similarity between the separated audio embedding and the text query ("the sound of tap dancing") embedding generated by the CLAP model.

Model	Text Query	SDR	SIR	SAR	CLAP score
CLIPSEP-NIT (Dong et al., 2023)		11.8540	24.1064	16.3873	0.3053
OmniSep (Cheng et al., 2025b)	"tap dancing"	10.8925	**24.2579**	17.0300	0.4810
MARS-Sep (Ours)		**12.0603**	24.0055	**17.2554**	**0.4935**

This case study validates that by directly optimizing for semantic consistency, MARS-Sep effectively mitigates semantic contamination and delivers perceptually superior results in scenarios where traditional signal-level metrics can be misleading. Moreover, the use of CLAP score highlights the advantage of employing cross-modal semantic evaluation, as it aligns closely with human perception of whether the separation captures the intended sound concept.

E.3 Efficacy of the Progressive Alignment Finetuning

To further examine the contribution of the progressive alignment strategy in our *MARS-Sep* framework, we replace the progressively fine-tuned ImageBind encoder with

(i) a frozen version without any fine-tuning (*no finetuning*),

(ii) a variant fine-tuned in a single stage on the mixed paired dataset (*1-stage finetuning*).

Table 6: Comparison of sound separation performance among different fine-tuning strategies on **VGGSOUND-clean+** dataset.

Methods	VGGSOUND-clean+				
	Mean SDR↑	Mean SIR↑	Mean SAR↑	Mean SI-SDRi↑	Mean $CLAP_t$ ↑
Text Query Sound Separation					
No fine-tuning	6.59±0.68	8.82±1.01	13.67±0.75	4.23±0.46	8.56±0.90
1-stage fine-tuning	6.73±0.68	9.24±0.99	13.72±0.79	4.40±0.46	**9.07±0.91**
3-stage fine-tuning	**6.91±0.68**	**9.14±1.00**	**13.73±0.77**	**4.55±0.44**	9.03±0.94
Audio Query Sound Separation					
No fine-tuning	6.85±0.62	11.46±0.99	11.39±0.77	4.15±0.53	8.69±0.93
1-stage fine-tuning	6.69±0.62	11.35±1.03	11.40±0.78	3.98±0.53	8.64±0.91
3-stage fine-tuning	**7.33±0.67**	**11.63±1.00**	**12.00±0.84**	**4.36±0.50**	**8.91±0.91**
Image Query Sound Separation					
No fine-tuning	**7.11±0.68**	9.96±1.04	**14.00±0.75**	**4.68±0.48**	8.59±0.91
1-stage fine-tuning	6.54±0.63	9.99±1.05	13.57±0.77	4.11±0.50	9.17±0.90
3-stage fine-tuning	6.93±0.67	**10.18±1.04**	13.41±0.72	4.57±0.47	**9.19±0.91**
Composed Omni-modal Query Sound Separation					
No fine-tuning	7.69±0.75	10.34±1.01	**14.57±0.94**	4.98±0.48	9.05±0.91
1-stage fine-tuning	7.67±0.74	**10.70±1.03**	14.48±0.94	4.84±0.50	8.83±0.89
3-stage fine-tuning	**7.93±0.75**	10.65±1.00	14.49±0.95	**5.20±0.45**	**9.22±0.90**

As shown by the results in Table 6, *MARS-Sep* with progressive alignment still demonstrates clear advantages on the **VGGSOUND-clean+** dataset, though the trends differ slightly from those observed on MUSIC-clean+. For text and audio query separation, the three-stage fine-tuning strategy

consistently yields the best overall signal-level metrics (SDR, SIR, SAR, and SI-SDRi), indicating that progressive, stage-wise alignment effectively enhances the multimodal encoder's domain adaptation capability. In particular, the gain in audio query separation is most evident, where the three-stage strategy surpasses both the frozen and single-stage variants across all metrics, showing its robustness in cross-modal grounding.

However, for image query separation, the improvement is less consistent - the no-fine-tuning model performs competitively on SDR and SAR, suggesting that visual embeddings pre-trained on large-scale video data are already well-aligned with audio representations in VGGSOUND-clean+. Nonetheless, fine-tuning slightly boosts the CLAP similarity, implying better semantic alignment between modalities.

When all modalities are jointly composed as omni-modal queries, the three-stage fine-tuning strategy again achieves the best balance between separation quality and semantic consistency, achieving the highest SI-SDRi and CLAP_t scores. Overall, these results confirm that while the benefits of progressive alignment may vary across modalities, it remains most effective for improving cross-domain generalization and semantic consistency in complex, in-the-wild sound mixtures.

E.4 Additional Qualitative Results in the TQSS Setting

Figure 4 illustrates representative qualitative comparisons under the TQSS scenario. For each example, we visualize the log-mel spectrograms of the mixed input, the target source, the interference source, as well as the separation outputs from the baseline method and our proposed approach. As can be observed, our method better preserves the structure of the target source while effectively suppressing interference components. More examples are available on our project webpage `https://mars-sep.github.io/`.

F The Use of Large Language Models (LLMs)

In this work, we utilized a large language model (GPT-5) for two auxiliary purposes. It was used to generate illustrative images, including those of a 'fox barking' and a 'playing congas' for Figure 2. Additionally, we leveraged its search capabilities to assist with our literature review.

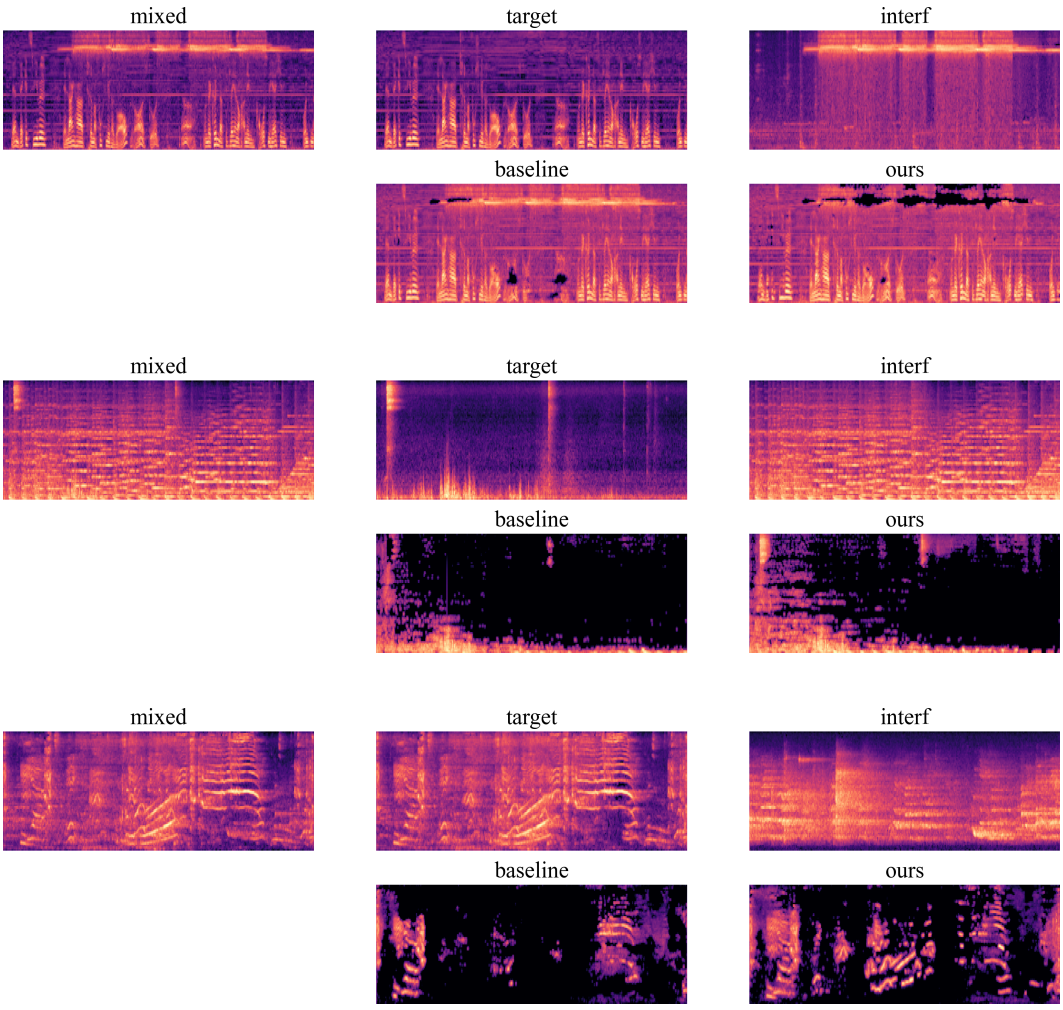

Figure 4: Qualitative comparison of separation results in the TQSS setting. Each group contains 5 spectrograms: mixed input, target source, interference source, baseline(OmniSep) separation, and our method separation.