VocalRep: Structure-Aware Vocal Representations for Multimodal Generation

Anonymous ACL submission

Abstract

Modern speech and multimodal generation systems, such as singing voice conversion and audio-driven lip synchronization, critically depend on temporally stable and semantically unambiguous vocal representations. In practical pipelines, such representations are typically derived from music source separation (MSS) applied to mixed musical recordings. However, standard MSS paradigms often aggregate lead vocals and backing harmonies into a single vocal stream. Although multi-stem separation has been explored, existing approaches remain primarily optimized for signal-level reconstruction, often overlooking the intricate structural disentanglement required by downstream generation tasks. From a generation-oriented perspective, this motivates revisiting vocal separation from a representation learning standpoint. To this end, we propose VocalRep, a structure-aware learning framework designed to disentangle lead vocals, harmonies, and instrument while enforcing role consistency across long-form audio. By integrating global vocal identity conditioning with ranking-based objectives, VocalRep extracts role-consistent lead vocal representations without relying on explicit pitch or symbolic annotations. Experimental results demonstrate that VocalRep significantly improves performance in downstream singing voice conversion and audio-driven lip synchronization.

1 Introduction

Recent advances in speech and multimodal generation have enabled increasingly powerful systems for singing voice conversion (SVC), singing voice synthesis, and audio-driven lip synchronization (Liu et al., 2021; Huang et al., 2023; Zheng et al., 2025; Suwajanakorn et al., 2017; Zhang et al., 2023). Despite their architectural diversity, these systems share a fundamental dependency on temporally stable and semantically coherent acoustic representations. In practical pipelines, such representations are often obtained from mixed musical recordings through music source separation (MSS) models (Défossez et al., 2019).

Most existing MSS approaches, however, are optimized using reconstruction-oriented objectives and evaluated with metrics such as SDR or SI-SDR (Rafii et al., 2017; Zhang et al., 2025). While effective for measuring signal-level similarity, these metrics are largely insensitive to human perceptual quality or the representational suitability for downstream tasks like sequence modeling, pitch inference, and phoneme-level alignment (Jaffe and Burgoyne, 2025; Cano et al., 2016). Consequently, separation results that achieve high conventional scores often harbor structural artifacts, unstable pitch trajectories, and reduced intelligibility, proving detrimental when employed in generative modeling.

A key source of this limitation lies in how contemporary MSS systems model vocal content. Most widely adopted approaches perform two-stem separation, aggregating lead vocals and harmonies into a single vocal stream (Rafii et al., 2019). Although this formulation simplifies optimization, it overlooks the distinct functional roles of lead vocals and harmonies in generation-oriented tasks. Specifically, SVC systems require representations that preserve a single, dominant pitch trajectory (Wei et al., 2023), whereas lip synchronization models depend on temporally unambiguous phonetic cues (McAuliffe et al., 2017). When multiple vocal roles are entangled within a single representation, downstream models are forced to operate on ambiguous inputs, resulting in pitch jitter, phoneme confusion, or temporal instability—even when separation metrics indicate strong performance.

Disentangling lead vocals from harmonies is particularly challenging because the problem is locally ambiguous. Locally, harmonies may share the lead's timbre, or—in the case of distinct back-

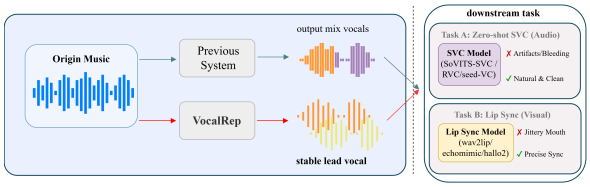

Figure 1: Comparison between conventional single-source separation and VocalRep. Role-aware vocal allocation yields more stable lead vocals for downstream tasks.

ing vocalists—momentarily assume melodic prominence during transitional segments, making them difficult to distinguish from the main vocal line. In contrast, vocal roles are typically well-defined at the song level, where the lead vocal exhibits long-term identity consistency and structural dominance. This mismatch between local role ambiguity and global structural consistency motivates the need for structure-aware representation learning.

To address these challenges, we propose VocalRep, a structure-aware learning framework designed to recover generation-friendly vocal representations from polyphonic music. VocalRep incorporates global vocal identity conditioning to enforce long-range role consistency, and introduces ranking-based task-aware and structure-aware objectives that encourage representations to preserve vocal–instrument coherence and lead–harmony dominance. These objectives are implemented via discriminative critics used only during training, without introducing additional inference cost.

We evaluate VocalRep not only on conventional separation benchmarks, but also through end-to-end downstream generation tasks, including multiple state-of-the-art singing voice conversion systems and audio-driven lip synchronization models. Experimental results demonstrate that, while VocalRep achieves competitive separation scores, its primary advantage lies in producing representations that significantly improve intelligibility, pitch stability, and perceptual quality for downstream models. Our findings suggest that downstream generation performance provides a more faithful measure of representation quality than reconstruction-oriented separation metrics alone.

The main contributions of this work are summarized as follows[1]:

- We revisit music source separation to prioritize semantic purity for downstream generation, advocating for a three-stem paradigm (lead, harmony, instrument) that explicitly isolates lead vocals to support high-fidelity voice conversion and lip-sync.
- We propose VocalRep, a unified framework utilizing global vocal identity and ranking objectives. This approach resolves local role ambiguity and models the hierarchical interplay between vocals and instrument, significantly enhancing separation robustness.
- We establish a comprehensive benchmark bridging separation metrics with downstream SVC and lip-sync performance, and will release our code and models to facilitate application-oriented research.

[1]Demos are available at https://anonymous.4open.science/r/VocalRep_-EC38.

2 Related Work

2.1 State-of-the-Art Music Source Separation

Driven by Sound Demixing Challenge (SDX) competitions, MSS has rapidly evolved from CNN-based U-Nets to Transformer-based architectures with explicit band modeling (Stöter et al., 2019; Hennequin et al., 2020). MDX23C achieves strong separation performance but relies on large-scale models and ensemble strategies (Kim et al., 2023; Solovyev et al., 2023). BS-Roformer (Lu et al., 2024) introduces the band-split (BS) module combined with the transformer to improve frequency-specific representations. MelBand-Roformer (Wang et al., 2023b) improves perceptual coherence by adopting overlapping Mel-scale band mappings. SCNet (Tong et al., 2024) leverages sparse complex-domain modeling to efficiently recover fine spectral details.

Despite architectural advances, the underlying separation assumption remains unchanged, existing methods largely treat vocals as a single spectral entity and fail to disentangle the internal polyphonic structure of human voices. Separating lead vocals from harmonies is a more challenging sub-task than general music separation. Most existing approaches rely on retraining general-purpose architectures such as U-Net–based or other on datasets with fine-grained annotations (Bittner et al., 2014; Jansson et al., 2017; Rouard et al., 2023; Bai et al., 2024). A few specialized studies exploit multi-channel or spatial cues, but such assumptions break down for single-channel studio mixes (Cano et al., 2018). As a result, current methods struggle to produce semantically clean lead vocals suitable for downstream generative tasks.

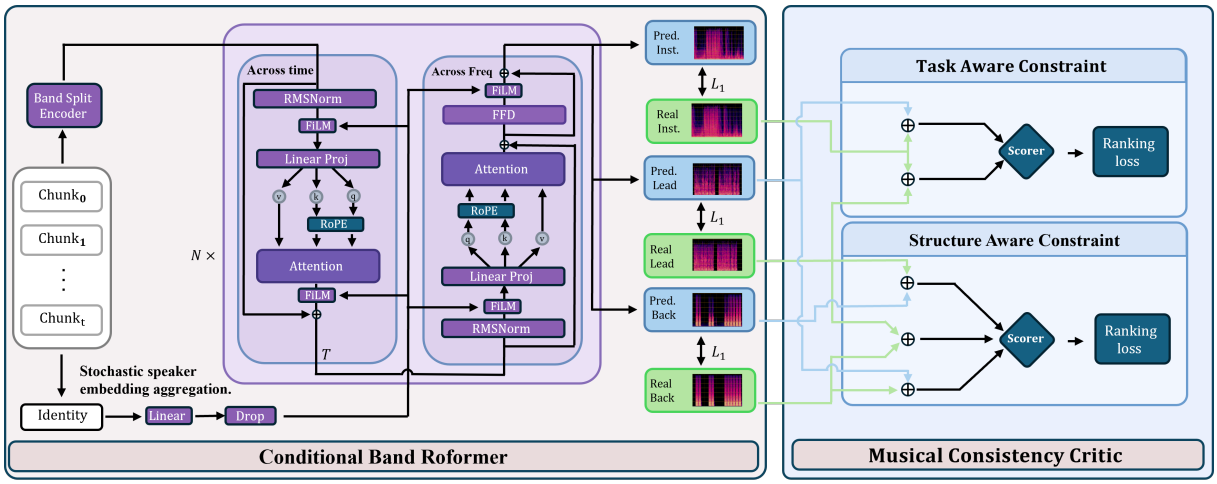

Figure 2: The overall architecture of VocalRep. The Conditional Roformer Backbone (Left) separates the input mixture into Lead, Harmony, and instrument stems. The Musical Consistency Critic (Right) enforces structural realism using Task-Aware and Structure-Aware constraints.

2.2 Downstream Generation Tasks

With the rapid advancement of multimodal generation AI, high-quality audio separation is no longer merely an end goal but a core pre-processing module serving downstream tasks. However, existing separation research is often disconnected from downstream applications, resulting in a significant misalignment between evaluation metrics (e.g., SDR) and practical usability (Chen et al., 2024).

Robustness Challenges in SVC. SVC systems rely heavily on precise F0 estimation and content feature extraction from the source audio (Kim et al., 2018). Research indicates that F0 extractors are extremely sensitive to non-harmonic interference and multi-voice aliasing. When the separation model fails to completely strip away harmonies, residual harmonic components induce pitch jitter or octave errors in the F0 estimation algorithm (Kawahara et al., 2001; Salamon and Gómez, 2012; Reghunath et al., 2025). These "signal-level" residuals are amplified by the SVC model, causing the generated singing voice to exhibit severe robotic artifacts or detuning, even if the separation results appear acceptable in terms of SDR (Kim et al., 2021).

High-Precision Lip Synchronization. In Lip Sync tasks, models are required to extract high-level phonemes or semantic representations from audio to drive facial landmarks (Prajwal et al., 2020; Zhang et al., 2023). Multi-voice mixtures blur the temporal boundaries of phonemes, leading to "ambiguous" feature inputs.

3 Methodology

As illustrated in Fig. 2, this section presents the proposed VocalRep framework for vocal separation in polyphonic music. The model adopts a conditional Band-Roformer as the primary source estimator, which decomposes the input mixture into lead vocals, harmonies, and instrument stems. To promote musically plausible separation beyond waveform reconstruction, the backbone is jointly optimized with two structural consistency constraints: a task-aware constraint that enforces extrinsic coherence between the lead vocal and its instrumental context, and a structure-aware constraint that captures the intrinsic harmonic coupling between lead vocals and harmonies.

3.1 Problem Formulation

Given a polyphonic musical recording $x \in \mathbb{R}^T$, our goal is to derive vocal representations that are suitable for speech and multimodal generation tasks. Instead of treating separation as a purely reconstruction-driven problem, we explicitly model three functional components: lead vocals s^{lead}, harmonies s^{harm}, and instrument s^{inst}. We learn a parameterized model F_ϕ that maps the mixture signal to component-wise waveform estimates:

$$(\hat{s}^{\text{lead}}, \hat{s}^{\text{harm}}, \hat{s}^{\text{inst}}) = F_\phi(x). \quad (1)$$

To ensure basic signal fidelity, we employ a standard reconstruction objective that encourages component-wise consistency with the reference tracks:

$$\mathcal{L}_{\text{sep}} = \sum_{r \in \{\text{lead,harm,inst}\}} \|\hat{s}^r - s^r\|_1. \quad (2)$$

While this objective promotes output similarity, it does not enforce role stability or structural coherence among vocal components. In the following, we detail the conditional backbone and the structure-aware perception mechanisms designed to address these limitations.

3.2 Conditional Band-Roformer Backbone

We employ a Band-Split Roformer conditioned on global vocal identity to resolve local role ambiguities. Given the input spectrogram $\mathbf{X}$, the encoder features $\mathbf{H} = \mathrm{Enc}(\mathbf{X})$ are processed by a Transformer conditioned on a global vector $\mathbf{c}$ via feature-wise linear modulation (FiLM) (Perez et al., 2018):

$$\mathbf{Z} = \mathrm{Transformer}(\mathbf{H}; \mathbf{c}). \tag{3}$$

The vector $\mathbf{c}$ is derived from a pre-trained CAM++ encoder (Yu et al., 2021; Wang et al., 2023a) using two distinct strategies:

Training-time aggregation. We extract embeddings from random segments of the ground-truth lead vocal. At each iteration, $N = 20$ embeddings are averaged to form $\mathbf{c}$. To prevent over-reliance on identity cues and enhance robustness, $\mathbf{c}$ is randomly dropped with a probability of 0.5.

Inference-time anchoring. Since ground-truth vocals are unavailable, we perform spectral clustering on embeddings extracted from sliding windows of the input mixture. The centroid of the dominant cluster is selected as the song-level anchor $\mathbf{c}$, ensuring consistent lead vocal identification throughout the track.

Finally, we employ three distinct prediction heads f_ϕ^r to generate component-specific masks and reconstruct waveforms via inverse STFT:

$$\hat{s}^r = \mathrm{iSTFT}\big(f_\phi^r(\mathbf{Z}) \odot \mathbf{X}\big), \tag{4}$$

where $r \in \{\mathrm{lead}, \mathrm{harm}, \mathrm{inst}\}$.

3.3 Task-Aware Network

Vanilla reconstruction objectives typically treat the lead vocal independently from the instrument, thereby failing to enforce **vocal-instrumental compatibility**. To address this limitation, we introduce a task-aware network that explicitly evaluates the rhythmic coherence of the separated lead vocal within its musical context. A theoretical interpretation of this ranking-based formulation, from an energy-based and symmetry-breaking perspective, is provided in Appendix A.

We define the natural reference mixture and the estimated mixture as

$$x_{\mathrm{ref}} = s^{\mathrm{lead}} + s^{\mathrm{inst}}, \tag{5}$$

$$x_{\mathrm{est}} = \hat{s}^{\mathrm{lead}} + s^{\mathrm{inst}}. \tag{6}$$

A task-aware critic P_ψ assigns plausibility scores to these mixtures and is trained to rank the reference higher than the estimated composition:

$$\mathcal{L}_{\mathrm{rank}}^{P} = \mathbb{E}[\mathrm{softplus}(P_\psi(x_{\mathrm{est}}) - P_\psi(x_{\mathrm{ref}}))]\,. \tag{7}$$

The separation backbone is optimized with the reversed objective, encouraging the predicted lead vocal to remain compatible with the ground-truth instrument:

$$\mathcal{L}_{\mathrm{rank}}^{S} = \mathbb{E}[\mathrm{softplus}(P_\psi(x_{\mathrm{ref}}) - P_\psi(x_{\mathrm{est}}))]\,. \tag{8}$$

Through this relative ranking formulation, the critic learns a shared scoring space in which mixtures with and without instrument, as well as configurations with and without a dominant lead vocal, are mapped to comparable plausibility scores.

3.4 Structure-Aware Network

Polyphonic vocals exhibit strong intrinsic coupling, where lead vocals and harmonies share similar timbral and melodic characteristics. To explicitly encourage role separation, we introduce a structure-aware network that enforces structural consistency through cross-composition.

We construct two validation mixtures,

$$x_{\mathrm{lh}}^{(1)} = \hat{s}^{\mathrm{lead}} + s^{\mathrm{harm}}, \tag{9}$$

$$x_{\mathrm{lh}}^{(2)} = s^{\mathrm{lead}} + \hat{s}^{\mathrm{harm}}. \tag{10}$$

and contrast them against the pristine vocal mixture $x_{\mathrm{vocal}} = s^{\mathrm{lead}} + s^{\mathrm{harm}}$. A structure-aware critic P_θ is trained to rank the natural vocal composition higher than the cross-composed ones:

$$\mathcal{L}_{\mathrm{rank}}^{\mathrm{vh},P} = \mathbb{E}\left[\sum_{i=1}^{2} \mathrm{softplus}\Big(P_\theta(x_{\mathrm{lh}}^{(i)}) - P_\theta(x_{\mathrm{vocal}})\Big)\right]. \tag{11}$$

The separation network minimizes the inverse objective, encouraging predictions that preserve the structural dominance of the lead vocal while suppressing harmonic interference:

$$\mathcal{L}_{\mathrm{rank}}^{\mathrm{vh},S} = \mathbb{E}\left[\sum_{i=1}^{2} \mathrm{softplus}\Big(P_\theta(x_{\mathrm{vocal}}) - P_\theta(x_{\mathrm{lh}}^{(i)})\Big)\right]. \tag{12}$$

Through this cross-composition ranking scheme, the critic learns a unified structural scoring space in which natural and role-swapped vocal configurations are mapped to comparable scores, enabling consistent discrimination of lead vocals and harmonies across different vocal mixtures.

3.5 Total Optimization Objective

The framework is trained using an alternating optimization strategy. The separation backbone F_ϕ minimizes a composite objective that unifies waveform reconstruction with structural consistency:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sep}} + \lambda_{\text{task}}\mathcal{L}^{S}_{\text{rank}} + \lambda_{\text{struct}}\mathcal{L}^{\text{vh},S}_{\text{rank}}, \quad (13)$$

where λ_{task} and λ_{struct} are hyperparameters balancing the auxiliary constraints. Concurrently, the critics P_ψ and P_θ are optimized to minimize their respective discrimination losses (Eq. 7 and Eq. 11), providing gradients to guide the separator toward generation-friendly representations.

4 Experimental Setup

4.1 Implementation Details

Training samples are segmented into 8-second clips at 44.1 kHz. We optimize the model using Adam on four NVIDIA A800 GPUs with a total batch size of 4. Both critics employ an identical ensemble of Multi-Period (MPD), Multi-Scale (MSD) (Kong et al., 2020), and Multi-Resolution Spectrogram (MR-Spec) (Jang et al., 2021) discriminators, ensuring comprehensive coverage of time-frequency patterns. These auxiliary networks are discarded during inference. Detailed configurations are available in Appendix B.2.

4.2 Datasets and Baselines

We train on a 250-hour in-house dataset and evaluate on three benchmarks: the public MUSDB18-HQ and Multisong (from MVSeq) sets, and our curated PHV-40 (Polyphonic Harmonies Vocals) dataset. For benchmarking, we compare VocalRep against a diverse suite of SOTA architectures, including Band-Split Roformers and Hybrid Transformers. To ensure rigor and reproducibility, we utilize widely adopted community checkpoints for all baselines. Detailed specifications for datasets and model configurations are provided in Appendix B.3 and Appendix B.4, respectively.

4.3 Evaluation Metrics

We employ a multi-dimensional evaluation protocol. For signal fidelity, we report standard SDR, SI-SDR, Bleedless (quantifying vocal cleanliness), and Fullness (measuring instrumental richness). To assess downstream utility, we evaluate Singing Voice Conversion (SVC) via CER (intelligibility), logF0-PCC (pitch consistency), SPK-SIM (identity), and UTMOS (aesthetics). Furthermore, Audio-Driven Lip Synchronization is evaluated using Sync-D (distance, lower is better) and Sync-C (confidence, higher is better). Please refer to Appendix C for detailed definitions and formulations.

5 Results and Discussions

5.1 Analysis on Source Separation

We first evaluate VocalRep on standard 2-stem benchmarks (MUSDB18-HQ and Multisong) as shown in Table 1. Our method achieves competitive performance, with an SI-SDR of 11.45 dB on MUSDB18-HQ and 9.69 dB on Multisong. We observe a slight performance gap compared to top-tier specialized models like Band-Roformer (11.85 dB). This is expected, as our reported vocal stem is a summation of separately predicted lead and harmony tracks, a process that naturally accumulates residual errors from both sub-tasks. However, this trade-off is acceptable given our model's unique capability to provide fine-grained control over vocal components.

On the harmonic-dense PHV-40 dataset, while our summed vocal scores are slightly lower than the baselines, we achieve a substantial improvement in instrument separation quality (16.35 dB vs. 10.60 dB for MelBand-Roformer). This indicates that by explicitly modeling the harmonic structure, VocalRep more effectively disentangles complex vocal textures from the instrument, reducing leakage even if the reconstruction loss for individual vocal stems is marginally higher.

5.2 Fine-Grained Harmonic Disentanglement

Table 2 presents the results on the 3-stem separation task (Lead / Harmony / Instrument), which is particularly challenging for standard separation models. For the harmony stem, several competitive baselines exhibit low or even negative SI-SDR values, indicating substantial interference between vocal components. Specifically, MelBand-Roformer and MDX23C achieve SI-SDR scores of -0.39 dB and -4.23 dB, respectively, while SCNet Large attains a positive SI-SDR of 1.65 dB. These results suggest that accurately separating harmonies from both lead vocals and instrument remains difficult

Table 1: Separation performance of different front-end models on three datasets. For our method, the separated lead and harmony stems are summed into a single vocal stem for comparison. The best result is **bolded**, the second best is underlined, and the third best is *italicized*.

System	Vocal			Instrument		
	SDR↑	SI-SDR↑	L1-Freq↓	SDR↑	SI-SDR↑	L1-Freq↓
Multisong						
MelBand-Roformer (Wang et al., 2023b)	**10.98**	**10.53**	38.87	**17.28**	**17.18**	39.68
Band-Roformer (Lu et al., 2024)	10.88	10.38	38.08	17.19	17.09	39.06
MDX23C (Kim et al., 2023)	10.15	9.59	35.91	16.46	16.34	*36.67*
SCNet Large (Tong et al., 2024)	9.06	8.24	**32.81**	–	–	–
HTDemucs4 (Rouard et al., 2023)	8.79	8.06	33.48	15.09	14.93	**34.36**
BS-Conformer	8.76	7.92	*35.18*	15.07	14.92	35.19
VocalRep (Ours)	*10.23*	*9.69*	36.08	*16.74*	*16.64*	37.99
MUSDB18-HQ†						
MelBand-Roformer (Wang et al., 2023b)	12.14	11.73	54.40	18.60	18.65	54.53
Band-Roformer (Lu et al., 2024)	**12.31**	**11.85**	53.98	**18.78**	**18.83**	54.21
MDX23C (Kim et al., 2023)	11.37	10.88	50.73	17.82	17.87	*50.60*
SCNet Large (Tong et al., 2024)	10.22	9.12	**45.89**	16.40	16.40	**42.54**
HTDemucs4 (Rouard et al., 2023)	9.33	8.51	47.64	15.66	15.65	47.21
BS-Conformer	11.05	10.60	*49.93*	17.09	16.99	51.37
VocalRep (Ours)	*11.78*	*11.45*	50.96	*18.38*	*18.31*	52.90
PHV-40 (Ours)						
MelBand-Roformer (Wang et al., 2023b)	9.98	13.27	28.40	11.43	10.60	31.33
Band-Roformer (Lu et al., 2024)	**10.08**	**13.49**	28.24	*11.54*	*10.78*	31.35
MDX23C (Kim et al., 2023)	9.64	11.59	27.30	**12.72**	15.91	**23.60**
SCNet Large (Tong et al., 2024)	9.28	12.33	**23.77**	10.76	9.91	*27.45*
HTDemucs4 (Rouard et al., 2023)	8.73	11.03	26.06	10.21	9.27	29.19
BS-Conformer	8.93	11.35	*26.30*	10.41	9.52	29.12
VocalRep (Ours)	*9.73*	*12.73*	27.54	11.85	**16.35**	24.39

† To avoid anomalous SI-SDR scores caused by low-energy normalization in local segments, and considering that loudness is encoded by the model in downstream tasks, we modified the validation pipeline to calculate scores on the entire song.

Table 2: Separation performance on the PHV-40. We explicitly report lead, harmony and instrument stems. The best result is **bolded** and the second best is underlined.

System	Lead			Harmony			Instrument		
	SDR↑	SI-SDR↑	Bleedless↑	SDR↑	SI-SDR↑	Bleedless↑	SDR↑	SI-SDR↑	Fullness↑
MelBand-Roformer†	6.67	6.59	18.95	2.72	-0.39	16.47	8.14	6.93	16.04
Band-Roformer†	6.94	**8.08**	18.06	3.49	0.78	**16.72**	11.30	10.32	17.57
MDX23C†	4.28	2.83	13.90	1.05	-4.23	12.38	10.34	9.08	**25.03**
SCNet Large†	6.31	6.51	16.52	3.48	1.65	15.39	8.25	7.07	18.58
VocalRep (Ours)	**7.19**	7.99	**19.96**	**4.02**	**2.87**	16.29	**11.76**	**15.70**	21.57

† Public models adapted to a 3-stem (lead / harmony / instrument) configuration and re-trained on our in-house dataset. Training is stopped if no improvement is observed for 50 epochs.

for conventional architectures.

In contrast, VocalRep achieves higher fidelity across all stems, with SI-SDR scores of 7.99 dB for lead vocals and 2.87 dB for harmonies. VocalRep also obtains a higher Bleedless score on the lead vocal stem, indicating cleaner separation with reduced cross-source interference. The most pronounced improvement is observed on the instrumental stem, where VocalRep achieves an SI-SDR of 15.70 dB, exceeding the strongest baseline (Band-Roformer, 10.32 dB) by more than 5 dB. This result indicates that improved vocal disentanglement leads to a cleaner instrumental estimate with reduced harmonic leakage.

The improvements indicate that fine-grained harmonic disentanglement reduces mutual interference between vocal components, benefiting both vocal and instrumental separation.

5.3 Downstream Evaluation on SVC

Table 3 evaluates three SVC systems using stems from different separation models. VocalRep consistently improves linguistic intelligibility across all backends. For instance, on Multisong with SoVITS-SVC, our method reduces the Character Error Rate (CER) to 36.49%, outperforming the strongest baseline (38.19%). With Seed-VC, VocalRep achieves 26.85% CER compared to 28.44% for the best competitor, indicating that removing harmonic interference effectively resolves phonetic ambiguity.

Beyond intelligibility, VocalRep excels in stabi-

Table 3: Performance of Singing Voice Conversion (SVC) systems on Multisong and PHV-40 datasets. We evaluate the quality of converted vocals using separated stems from different front-end models. The best result is **bolded** and the second best is underlined.

System	Multisong					PHV-40				
	SPK-SIM↑	CER(%)↓	Aesthetics		logF0PCC(%)↑	SPK-SIM↑	CER(%)↓	Aesthetics		logF0PCC(%)↑
			CE↑	CU↑				CE↑	CU↑	
SoVITS-SVC (svc-develop-team, 2023)										
MelBand-Roformer	0.604	41.80	5.63	6.17	54.25	**0.625**	43.09	5.56	6.21	42.04
+ stem-reconfigured	**0.616**	38.94	5.64	6.15	76.13	0.604	45.60	5.54	6.30	73.80
Band-Roformer	0.603	46.51	5.74	6.21	54.49	0.616	45.04	5.50	6.14	41.83
+ stem-reconfigured	0.596	38.19	**5.86**	**6.30**	79.89	0.590	40.40	5.45	6.14	73.04
MDX23C	0.591	43.71	5.65	6.03	55.45	0.609	45.31	5.47	6.05	42.36
+ stem-reconfigured	0.568	53.28	5.10	5.52	75.93	0.599	60.08	4.96	5.61	71.45
SCNet Large	0.582	42.93	5.60	5.99	79.41	0.608	60.53	5.37	5.89	43.93
+ stem-reconfigured	0.588	45.41	5.73	6.06	59.92	0.613	55.03	5.44	5.93	43.89
VocalRep (Ours)	0.595	**36.49**	5.81	6.13	**83.73**	0.581	**39.32**	**5.76**	**6.46**	**78.32**
RVC (RVC-Project, 2023)										
MelBand-Roformer	0.774	48.05	5.74	6.31	88.85	0.775	48.87	5.68	6.35	82.08
+ stem-reconfigured	0.767	42.72	5.74	6.36	90.79	0.769	47.72	5.57	6.38	89.62
Band-Roformer	0.773	49.60	5.75	6.32	88.39	0.773	51.73	5.68	6.36	81.68
+ stem-reconfigured	0.770	42.96	5.78	6.38	93.15	0.769	46.44	5.69	6.43	88.15
MDX23C	**0.775**	48.07	5.63	6.21	88.18	0.774	49.55	5.58	6.26	82.75
+ stem-reconfigured	0.751	57.91	5.08	5.85	86.40	0.755	59.93	4.96	5.84	81.91
SCNet Large	0.765	47.64	5.56	6.24	90.89	**0.777**	52.97	5.39	6.10	80.30
+ stem-reconfigured	0.774	53.26	5.48	6.08	87.64	**0.777**	51.06	5.39	6.10	80.36
VocalRep (Ours)	0.770	**39.80**	**5.85**	**6.43**	**95.15**	0.768	**43.58**	**5.73**	**6.51**	**90.81**
Seed-VC (Liu, 2024)										
MelBand-Roformer	0.825	32.22	5.66	6.24	87.24	0.821	29.16	**5.48**	6.28	80.50
+ stem-reconfigured	0.820	29.06	5.62	6.29	89.03	0.812	32.89	5.33	6.28	86.34
Band-Roformer	0.826	31.20	5.67	6.24	86.69	**0.822**	29.97	5.46	6.27	80.01
+ stem-reconfigured	0.824	28.44	5.70	6.31	91.14	0.818	**28.83**	5.45	6.31	85.60
MDX23C	**0.827**	32.55	5.60	6.16	86.45	0.821	31.08	5.44	6.19	81.04
+ stem-reconfigured	0.817	38.45	5.09	5.91	83.90	0.817	42.52	4.96	5.95	80.44
SCNet Large	0.822	31.23	5.48	6.13	87.85	0.820	33.02	5.37	6.17	78.61
+ stem-reconfigured	0.825	32.36	5.50	6.09	85.78	0.821	32.71	5.38	6.18	78.74
VocalRep (Ours)	0.824	**26.85**	**5.77**	**6.37**	**92.97**	0.815	29.15	5.46	**6.34**	**88.12**

lizing pitch trajectories, consistently achieving the highest logF0PCC scores. On the RVC backend, it reaches 95.15% on Multisong and 90.81% on PHV-40, significantly surpassing other front-ends. This suggests that suppressing residual backing vocals prevents incorrect pitch tracking, ensuring accurate melody preservation.

Crucially, this aggressive interference removal does not compromise audio quality or identity. VocalRep maintains competitive Speaker Similarity (SPK-SIM) scores—comparable to MelBand-Roformer and Band-Roformer (e.g., 0.770 on RVC)—and scores highest in perceptual aesthetics. This confirms VocalRep successfully disentangles lead vocal structure while preserving the naturalness required for high-fidelity voice conversion.

Overall, these results demonstrate that the benefits of VocalRep are not tied to a specific SVC backend, but generalize across different conversion architectures and datasets. By providing a cleaner and more structurally consistent lead-vocal representation, VocalRep improves downstream intelligibility and pitch accuracy without sacrificing speaker identity or perceptual quality.

5.4 Downstream Evaluation on Audio-Driven Lip Synchronization

Table 4 compares the performance of Wav2Lip, EchoMimic, and Hallo2 using vocal stems from different front-end models. VocalRep consistently outperforms baseline methods across all three lip-sync systems, demonstrating that cleaner vocal inputs translate directly to more accurate lip movements.

On the standard Wav2Lip benchmark, VocalRep achieves the lowest Sync-D scores (6.901 on Multisong and 6.850 on PHV-40), indicating superior synchronization accuracy. Furthermore, our method shows remarkable robustness on generative backends. With Hallo2, VocalRep attains the best performance in both distance (Sync-D) and confidence (Sync-C) metrics on both datasets (e.g., Sync-D 10.031 on PHV-40). By effectively removing interfering harmonies—which can trigger incorrect lip motion—VocalRep provides the most reliable phonetic guidance for high-fidelity talking head generation.

It is worth noting that the absolute performance metrics for the audio-driven digital human task are constrained by the nature of the evaluation data.

Table 4: Comparison of Audio-Driven Lip Synchronization systems (Wav2Lip, EchoMimic, Hallo2) across Multisong and PHV-40 datasets. We evaluate the synchronization quality using separated stems from different front-end models. Lower **Sync-D** indicates better lip-sync error, while higher **Sync-C** indicates better confidence.

Front-end Model	Wav2Lip (Prajwal et al., 2020)				EchoMimic (Chen et al., 2025)				Hallo2 (Cui et al., 2024)			
	Multisong		PHV-40		Multisong		PHV-40		Multisong		PHV-40	
	D↓	C↑	D↓	C↑	D↓	C↑	D↓	C↑	D↓	C↑	D↓	C↑
MelBand-Roformer	7.291	4.086	6.912	3.567	9.280	2.969	9.360	2.233	10.607	1.258	10.714	1.167
+ stem-reconfigured	7.042	4.104	7.028	3.272	8.984	2.519	9.058	2.547	10.368	1.132	10.371	1.134
Band-Roformer	7.323	4.061	6.931	3.540	9.449	3.092	**8.855**	2.978	10.368	1.503	10.413	1.182
+ stem-reconfigured	7.096	4.026	7.155	3.240	9.332	**3.160**	9.155	2.273	10.450	1.512	10.338	1.091
MDX23C	7.037	4.098	6.911	3.476	9.401	2.995	9.059	**3.100**	10.206	1.501	10.264	1.352
+ stem-reconfigured	7.687	3.472	7.302	2.849	8.792	2.182	9.680	2.047	10.361	0.893	10.302	0.976
SCNet Large	7.115	4.385	7.081	**3.714**	9.286	3.084	8.988	2.658	10.119	1.505	10.566	1.253
+ stem-reconfigured	7.412	4.141	7.149	3.160	9.294	2.336	9.057	2.760	10.349	1.134	10.119	1.227
VocalRep (Ours)	**6.901**	**4.391**	**6.850**	3.711	**8.725**	2.986	8.858	2.819	**10.110**	**1.577**	**10.031**	**1.404**

Table 5: Ablation study of individual components in **VocalRep** on the PHV-40. "ID" denotes explicit speaker identity–based correction applied during inference. Bold and underlined values indicate the best and second-best results within each column, respectively.

Configuration	Three-Stem SDR (dB)↑			RVC Performance				
	Lead	Harmony	Inst.	SPK-SIM↑	CER(%)↓	CE↑	CU↑	F0PCC(%)↑
Base separation backbone	6.94	3.49	11.30	0.769	46.44	5.69	6.43	88.15
+ Dual-Scorer Critics	7.24	4.02	11.16	0.770	**41.95**	**5.76**	6.49	90.00
+ Speaker Conditioning (inference w/o ID)	7.23	4.09	**11.80**	**0.774**	48.18	5.68	6.36	83.42
+ Speaker Conditioning (inference w/ ID)	**7.30**	**4.16**	11.78	0.769	43.90	5.71	6.47	89.95
+ Dual-Scorer + Speaker Cond. (inference w/o ID)	7.19	4.02	11.76	0.769	43.27	5.72	6.50	90.02
+ Dual-Scorer + Speaker Cond. (inference w/ ID)	7.29	4.05	11.75	0.768	43.58	5.73	**6.51**	**90.81**

This is primarily because the ground-truth videos contain polyphonic audio with multiple harmonic noises, whereas our approach produces cleaner vocal separation. This difference in audio quality leads to an inevitable discrepancy between the generated results and the ground-truth features. Nevertheless, compared with other methods, VocalRep yields the best relative performance on both Sync-D and Sync-C metrics, confirming its superiority in handling complex vocal scenarios.

5.5 Ablation Study and Analysis

Table 5 shows that both the dual-scorer critics and speaker conditioning contribute to improved separation and downstream RVC performance. The dual-scorer critics consistently improve lead and harmony SDR and reduce CER, indicating more effective vocal role disentanglement. Combining dual-scorer critics with speaker conditioning yields the best overall performance, demonstrating their complementary effects.

6 Conclusion

In this work, we revisit the task of vocal separation through the lens of downstream generation requirements. We identify vocal role ambiguity—specifically the entanglement of lead vocals and harmonies—as a critical factor that compromises performance in tasks such as singing voice conversion and audio-driven lip synchronization, even when conventional separation metrics indicate high performance. To address this, we propose VocalRep, a structure-aware learning framework designed to recover generation-oriented vocal representations from polyphonic music. By integrating global vocal identity conditioning with ranking-based structure-aware objectives, VocalRep enforces long-range role consistency and structural dominance without relying on explicit pitch extraction or symbolic supervision. Experimental results demonstrate that, while VocalRep achieves competitive scores on standard separation benchmarks, its primary advantage lies in substantially enhancing downstream generation quality, yielding superior intelligibility, pitch stability, and perceptual naturalness. These findings suggest that downstream generative performance serves as a more faithful proxy for representation quality than reconstruction-oriented metrics alone, highlighting the necessity of representation-aware modeling for future speech and multimodal generation systems.

Limitations

First, on standard 2-stem separation benchmarks, VocalRep scores slightly lower than top-tier specialized models. This is primarily attributed to our "separate-then-sum" strategy, this process inevitably leads to the accumulation of prediction errors from both sub-tasks. Second, the complexity of musical compositions poses a challenge. For songs that feature multiple lead vocalists (e.g., duets), the assumption of a single global speaker identity may be ill-defined. Future work will focus on addressing these limitations.

Ethics Statement

This work studies music source separation and vocal representation learning for downstream generation tasks such as singing voice conversion and audio-driven lip synchronization. The experiments are conducted on both publicly available benchmarks and a legally authorized in-house dataset. We strictly adhere to the licensing agreements and usage policies for all data sources. The proposed method does not introduce new data collection of personally identifiable information (PII) or user profiling. We acknowledge the potential risks of high-fidelity vocal extraction being misused for unauthorized voice cloning and advocate for responsible use and adherence to applicable ethical and legal standards.

References

Ye Bai, Chenxing Li, Hao Li, Yuanyuan Zhao, and Xiaorui Wang. 2024. Jointly recognizing speech and singing voices based on multi-task audio source separation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Rachel M Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. 2014. Medleydb: A multitrack dataset for annotation-intensive mir research. In *Ismir*, volume 14, pages 155–160.

Estefanía Cano, Derry FitzGerald, and Karlheinz Brandenburg. 2016. Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1758–1762. IEEE.

Estefania Cano, Derry FitzGerald, Antoine Liutkus, Mark D Plumbley, and Fabian-Robert Stöter. 2018. Musical source separation: An introduction. *IEEE Signal Processing Magazine*, 36(1):31–40.

Junyu Chen, Susmitha Vekkot, and Pancham Shukla. 2024. Music source separation based on a lightweight deep learning framework (dttnet: Dual-path tfc-tdf unet). In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 656–660. IEEE.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. 2025. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2403–2410.

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. 2024. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*.

Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. 2019. Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*.

Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154.

Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, and Tomoki Toda. 2023. The singing voice conversion challenge 2023. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Noah Jaffe and John Ashley Burgoyne. 2025. Musical source separation bake-off: Comparing objective metrics with human perception. *arXiv preprint arXiv:2507.06917*.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*.

Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. 2017. Singing voice separation with deep u-net convolutional networks.

Hideki Kawahara, Jo Estill, and Osamu Fujimura. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, pages 59–64. Firenze.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 161–165. IEEE.

Minseok Kim, Jun Hyung Lee, and Soonyoung Jung. 2023. Sound demixing challenge 2023 music demixing track technical report: Tfc-tdf-unet v3. *arXiv preprint arXiv:2306.09382*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Songting Liu. 2024. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*.

Songxiang Liu, Yuewen Cao, Dan Su, and Helen Meng. 2021. Diffsvc: A diffusion probabilistic model for singing voice conversion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 741–748. IEEE.

Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. 2024. Music source separation with band-split rope transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2017. The musdb18 corpus for music separation.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. 2019. Musdb18-hq-an uncompressed version of musdb18.

Lekshmi Chandrika Reghunath, Rajeev Rajan, Christian Napoli, and Cristian Randieri. 2025. Cross-attentive cnns for joint specral and pitch feature learning in predominant instrument recognition from polyphonic music. *Technologies*, 14(1):3.

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

RVC-Project. 2023. Retrieval-based voice conversion webui. https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI. GitHub repository, accessed 2025-01-06.

Justin Salamon and Emilia Gómez. 2012. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE transactions on audio, speech, and language processing*, 20(6):1759–1770.

Roman Solovyev, Alexander Stempkovskiy, and Tatiana Habruseva. 2023. Benchmarks and leaderboards for sound demixing tasks. *arXiv preprint arXiv:2305.07489*.

Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji. 2019. Open-unmix-a reference implementation for music source separation. *Journal of Open Source Software*, 4(41):1667.

Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13.

svc-develop-team. 2023. so-vits-svc: Softvc + vits singing voice conversion framework. https://github.com/svc-develop-team/so-vits-svc. GitHub repository, accessed 2025-01-06.

Weinan Tong, Jiaxu Zhu, Jun Chen, Shiyin Kang, Tao Jiang, Yang Li, Zhiyong Wu, and Helen Meng. 2024. Scnet: sparse compression network for music source separation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1276–1280. IEEE.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023a. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.

Ju-Chiang Wang, Wei-Tsung Lu, and Minz Won. 2023b. Mel-band roformer for music source separation. *arXiv preprint arXiv:2310.01809*.

Haojie Wei, Xueke Cao, Tangpeng Dan, and Yueguo Chen. 2023. Rmvpe: A robust model for vocal pitch estimation in polyphonic music. *arXiv preprint arXiv:2306.15412*.

Ya-Qi Yu, Siqi Zheng, Hongbin Suo, Yun Lei, and Wu-Jun Li. 2021. Cam: Context-aware masking for robust speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6703–6707. IEEE.

Chongbin Zhang, Jiaxiang Zheng, and Moxi Cao. 2025. A music source separation method integrating time–frequency decoupling and mamba-based state space modeling. *Scientific Reports*, 15(1):36280.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. 2023. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661.

Junjie Zheng, Gongyu Chen, Chaofan Ding, and Zihao Chen. 2025. R2-svc: Towards real-world robust and expressive zero-shot singing voice conversion. *arXiv preprint arXiv:2510.20677*.

A Theoretical Analysis: Symmetry Breaking via Ranking-Based Energy Modeling

This section provides an intuitive interpretation of lead–harmony confusion from a symmetry perspective and explains how the proposed ranking objective encourages consistent role assignment. Our analysis is framed using an energy-based view, without introducing additional modeling assumptions.

A.1 Symmetry in Lead–Harmony Separation

Standard supervised separation methods optimize a posterior of the form

$$p(\mathcal{S} \mid x) \propto p(x \mid \mathcal{S})\, p(\mathcal{S}), \tag{14}$$

where $\mathcal{S} = \{s^{\text{lead}}, s^{\text{harm}}, s^{\text{inst}}\}$. A common assumption is that the prior factorizes across sources.

In real music recordings, lead vocals and harmonies often share the same singer identity and exhibit similar timbral and melodic characteristics. As a result, their marginal distributions can be highly overlapping, i.e.,

$$p(s^{\text{lead}}) \approx p(s^{\text{harm}}). \tag{15}$$

Under this condition, swapping the roles of lead and harmony may yield comparable reconstruction quality. Consequently, reconstruction-based objectives are approximately invariant to role permutation, which leads to unstable role assignment in practice (e.g., role drifting across segments).

A.2 Energy-Based View with Asymmetric Role Constraint

To encourage consistent role assignment, we adopt an energy-based perspective and characterize the compatibility of vocal components through an energy function conditioned on the instrument:

$$E_\theta = E_\theta\big(s^{\text{lead}}, s^{\text{harm}}, s^{\text{inst}}\big). \tag{16}$$

Under this formulation, lower energy corresponds to more plausible role assignment, while higher energy indicates structural inconsistency. The associated conditional distribution can be implicitly defined as

$$p(s^{\text{lead}}, s^{\text{harm}} \mid s^{\text{inst}}) \;\propto\; \exp\big(-E_\theta\big). \tag{17}$$

We decompose the energy into a symmetric content term and an asymmetric role-related term:

$$\begin{aligned} E_\theta = &\underbrace{E_{\text{sym}}(s^{\text{lead}}) + E_{\text{sym}}(s^{\text{harm}})}_{\text{content fidelity}} \\ &+ \underbrace{\mathcal{E}_{\text{assign}}\big(s^{\text{lead}}, s^{\text{harm}} \mid s^{\text{inst}}\big)}_{\text{role consistency}}. \end{aligned} \tag{18}$$

The assignment term is intentionally asymmetric and penalizes configurations where the harmonic component dominates the lead vocal. Consistent separation therefore corresponds to lower energy for the correct role assignment than for its role-swapped counterpart.

A.3 Ranking Loss as a Practical Approximation

In practice, the assignment energy cannot be computed explicitly. Instead, we approximate relative energy differences using discriminators that score different vocal compositions. These scores can be viewed as proxies for negative energy.

The ranking loss encourages the score of a role-consistent composition to be higher than that of a role-swapped one:

$$\mathcal{L}_{\text{rank}} = \max\Big(0,\, \xi - \big[D(\mathcal{S}_{\text{real}}) - D(\mathcal{S}_{\text{swap}})\big]\Big), \tag{19}$$

where ξ is a margin. Minimizing this loss biases the optimization toward stable and consistent lead–harmony assignment. In our implementation, we initially set $\xi = 1$ to enforce a strict margin during the early training phase, and subsequently relax it to $\xi = 0$ once the model reaches stable convergence.

A.4 Implication: Implicit Lead Vocal Consistency

Under this formulation, explicit pitch or melody extraction is not required. Instead, lead vocal identification is guided implicitly by the ranking criteria.

At inference time, the predicted lead vocal can be interpreted as the component that best satisfies both contextual compatibility with the instrument and relative dominance over harmonies:

$$\begin{aligned} \hat{s}^{\text{lead}} = \arg\max_{s} \Big(&\mathbb{E}\big[\text{Coh}\big(s, s^{\text{inst}}\big)\big] \\ &+ \mathbb{E}\Big[\text{Dom}\big(s, s^{\text{harm}}\big)\Big] \Big). \end{aligned} \tag{20}$$

This explains how consistent lead vocal separation emerges from the training objective itself, without relying on explicit symbolic constraints.

B Experimental Setup

B.1 Evaluation Scope and Task Mapping

We provide a comprehensive breakdown of the evaluation tasks conducted across the three test datasets. The evaluation scope expands beyond standard source separation to include downstream generative tasks, specifically Singing Voice Conversion (SVC) and Lip Sync, verifying the practical utility of the separated stems.

Table 6: Mapping of datasets to specific evaluation tasks. PHV-40 serves as the primary benchmark for fine-grained disentanglement, while Multisong is utilized to validate downstream application performance.

Evaluation Task	MUSDB18-HQ	Multisong	PHV-40 (Ours)
Vocal-Accompaniment Separation	✓	✓	✓
Lead-Harmony Disentanglement	-	-	✓
SVC (Singing Voice Conversion)	-	✓	✓
Lip Sync Generation	-	✓	✓

B.2 Model Architecture and Configuration

The backbone of VocalRep is built upon the conditional Band-Split Roformer architecture, chosen for its superior capacity in modeling frequency-dependent dependencies. This backbone is configured to predict a three-stem output (lead vocals, harmonies, and instrument) consistent with our structure-aware paradigm. Detailed hyperparameters regarding the model architecture and training optimization are summarized in Table 7.

B.3 Dataset Specifications

MUSDB18-HQ: The standard industrial benchmark for music source separation. Since it lacks independent harmony stems, it is strictly used to evaluate the global quality of vocal-accompaniment separation.

Multisong (MVSeq): A public dataset containing polyphonic arrangements. In this work, we utilize it primarily to assess the distinctness and usability of separated components for downstream generative tasks, specifically **Singing Voice Conversion (SVC)** and **Lip Sync**, rather than for direct signal-level disentanglement metrics.

PHV-40: Our curated dataset consisting of 40 tracks with high-density vocal arrangements. It is the only dataset in our suite capable of supporting the full evaluation pipeline: from fine-grained Lead-Harmony Disentanglement to advanced downstream applications. Its complex overlapping vocals provide a rigorous testbed for measuring both structural separation accuracy and the fidelity of subsequent generation.

Table 7: Hyperparameters and training configuration of the VocalRep backbone.

Parameter	Value / Setting
Model Architecture	
Model Dimension	512
Transformer Depth	12 layers
Attention Heads	8 heads
Frequency Bands	Progressive (2 to 128 bins)
Dropout (Attn / FFN)	0.1 / 0.1
STFT (n_{fft}, hop, win)	2048, 441, 2048
Optimization Configuration	
Optimizer	Adam
Learning Rate	1×10^{-5}
Steps per Epoch	1000
Task Weight (λ_{task})	0.1
Struct Weight (λ_{struct})	0.1

B.4 Baseline Implementation Details

We provide the detailed list of baseline systems used in our evaluation in Table 9. To ensure a fair comparison representing the upper bound of current open-source performance, we selected high-performing checkpoints that are widely recognized in the MSS community (e.g., from the Ultimate Vocal Remover (UVR) ecosystem and establishing challenges).

B.5 SVC System Configurations

We detail the configurations of the Singing Voice Conversion (SVC) systems used in our downstream evaluation in Table 10. To assess the impact of stem quality across different conversion paradigms, we selected three representative frameworks: SoVITS-SVC (SoftVC VITS), RVC (retrieval-based VC), and Seed-VC (diffusion/flow-based). All systems employ RMVPE for pitch extraction and F_0 prediction.

B.6 Lip Synchronization Model Configurations

Table 11 summarizes the configurations of the audio-driven lip synchronization models used in our downstream evaluation. We select three widely adopted systems covering representative generation paradigms, including Wav2Lip, EchoMimic, and Hallo2. Wav2Lip follows a GAN-based formulation with explicit audio–visual alignment constraints, whereas EchoMimic and Hallo2 adopt latent diffusion frameworks with reference conditioning and long-form temporal modeling, respectively.

Table 8: Configurations and training hyperparameters of the music-friendly discriminators.

Configuration	Multi-Period (MPD)	Multi-Scale (MSD)	MR-Spectrogram (MR-Spec)
Input Domain	Waveform	Waveform	Magnitude Spectrogram
Core Mechanism	Period-based reshaping ($T \to T/p \times p$)	Multi-scale pooling (Temporal)	Multi-resolution STFT (Spectral)
Key Parameters	Periods: $\{2, 3, 5, 7, 11, 13\}$	Scales: $\{1, 2, 4, 8\}$	STFT Sizes: $\{1024, 2048, 4096\}$
Conv Type	2D Convolution	1D Convolution	2D Convolution
Network Depth	5 blocks	6 blocks per scale	5 blocks per resolution
Training Settings	Optimizer: Adam,	Learning Rate: 1×10^{-5},	Start Epoch: 100

Table 9: Detailed list of comparison models and their specific checkpoints/sources.

System	Source / Checkpoint Reference	URL
MelBand-Roformer	Weights by KimberleyJensen	https://github.com/KimberleyJensen/Mel-Band-Roformer-Vocal-Model
Band-Roformer	Weights by viperx	https://github.com/playdasegunda/band-split-rope-transformer
MDX23C	TFC-TDF-v3 (UVR Community)	https://github.com/ZFTurbo/MVSEP-MDX23-music-separation-model
SCNet Large	Implementation by starrytong	https://github.com/starrytong/SCNet
HTDemucs4	Meta AI (Fine-tuned by MVSep)	https://github.com/ZFTurbo/Music-Source-Separation-Training
BS-Conformer	Implementation by ZFTurbo	https://github.com/ZFTurbo/Music-Source-Separation-Training

C Detailed Definitions of Evaluation Metrics

This section provides formal definitions of the evaluation metrics used in our experiments. All metrics are computed on time-aligned signals after truncating them to the same length.

C.1 Multi-Source Separation Metrics

Signal Fidelity Metrics. Let $\mathbf{x} \in \mathbb{R}^{C \times T}$ denote the reference signal and $\hat{\mathbf{x}} \in \mathbb{R}^{C \times T}$ the estimated signal, where C is the number of channels and T is the number of samples.

Signal-to-Distortion Ratio (SDR). SDR measures overall reconstruction fidelity and is defined as

$$\mathrm{SDR} = 10 \log_{10} \frac{\|\mathbf{x}\|_2^2}{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \epsilon}, \tag{21}$$

where ϵ is a small constant for numerical stability. Higher SDR indicates better reconstruction fidelity.

Scale-Invariant SDR (SI-SDR). SI-SDR removes sensitivity to global amplitude scaling by projecting the estimate onto the reference signal:

$$\alpha = \frac{\langle \hat{\mathbf{x}}, \mathbf{x} \rangle}{\|\mathbf{x}\|_2^2 + \epsilon}, \tag{22}$$

$$\text{SI-SDR} = 10 \log_{10} \frac{\|\alpha \mathbf{x}\|_2^2}{\|\alpha \mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \epsilon}. \tag{23}$$

Higher SI-SDR indicates better scale-invariant waveform similarity.

Spectral Consistency Metrics. Let $|\mathcal{S}(\cdot)|$ denote the magnitude of the short-time Fourier transform (STFT).

L1 Frequency Metric (L1Freq). L1Freq computes the L_1 distance between magnitude spectrograms:

$$\mathcal{L}_{\mathrm{L1}} = \left\| |\mathcal{S}(\hat{\mathbf{x}})| - |\mathcal{S}(\mathbf{x})| \right\|_1. \tag{24}$$

The loss is mapped to a bounded score:

$$\mathrm{L1Freq} = \frac{100}{1 + \lambda \mathcal{L}_{\mathrm{L1}}}, \tag{25}$$

where λ is a scaling constant. Higher values indicate closer spectral alignment.

Source Separation Quality Metrics. Let $\mathrm{M}(\cdot)$ denote the mel-spectrogram in the logarithmic amplitude scale.

Bleedless and Fullness. Both metrics are computed from the mel-spectrogram difference:

$$\Delta = \mathrm{M}(\hat{\mathbf{x}}) - \mathrm{M}(\mathbf{x}). \tag{26}$$

Bleedless penalizes excessive positive deviations (energy leakage):

$$\mathrm{Bleedless} = \frac{100}{1 + \mathbb{E}[\Delta \mid \Delta > 0]}, \tag{27}$$

while Fullness penalizes excessive negative deviations (energy loss):

$$\mathrm{Fullness} = \frac{100}{1 + \mathbb{E}[-\Delta \mid \Delta < 0]}. \tag{28}$$

Higher values indicate cleaner separation and better energy preservation, respectively.

Table 10: Configurations and sources of the Singing Voice Conversion (SVC) systems.

System	Backbone / Key Components	Source / URL
SoVITS-SVC	VITS, Soft-VC Encoder: ContentVec	https://github.com/svc-develop-team/so-vits-svc
RVC	VITS, Faiss Retrieval Encoder: HuBERT-Soft	https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI
Seed-VC	DiT, Flow Matching Encoder: OpenAI Whisper	https://github.com/Plachtaa/Seed-VC

Table 11: Configurations and sources of the audio-driven lip synchronization systems used in our experiments.

System	Backbone / Paradigm	Source / URL
Wav2Lip	CNN-based Generator with Visual Discriminator SyncNet-based Lip–Audio Synchronization Loss	https://github.com/instant-high/wav2lip-onnx-256
EchoMimic	Latent Diffusion with Reference Conditioning UNet-based Diffusion Backbone	https://github.com/antgroup/echomimic
Hallo2	Audio-conditioned Diffusion with Transformer-based Conditioning Hierarchical Audio Encoding	https://huggingface.co/fudan-generative-ai/hallo2

C.2 Downstream Task Metrics

Singing Voice Conversion (SVC) Metrics. Given a converted waveform $\hat{y}$ and its reference (target) waveform y, we evaluate SVC performance from four complementary aspects.

Character Error Rate (CER). CER measures linguistic intelligibility by comparing the recognized text from $\hat{y}$ with the reference transcript. Let $\mathcal{T}(\cdot)$ denote a pretrained ASR system. In our experiments, we use *readASR* to obtain character sequences $\hat{\mathbf{s}} = \mathcal{T}(\hat{y})$ and $\mathbf{s} = \mathcal{T}(y)$, and compute the normalized edit distance:

$$\mathrm{CER} = \frac{D(\hat{\mathbf{s}}, \mathbf{s})}{|\mathbf{s}|} \times 100\%, \tag{29}$$

where $D(\cdot,\cdot)$ denotes the Levenshtein distance. Lower CER indicates better intelligibility.

logF0-PCC. logF0-PCC evaluates pitch contour consistency between the converted signal and the reference. Let $f_0(t)$ and $\hat{f}_0(t)$ denote frame-level fundamental frequency (F0) tracks extracted from y and $\hat{y}$, respectively. For fair comparison, we use the same pitch extractor (RMVPE) for all systems. The Pearson correlation coefficient is computed in the logarithmic domain over voiced frames $\mathcal{V}$:

$$\text{logF0-PCC} = \mathrm{corr}\Big(\log f_0(t), \log \hat{f}_0(t)\Big)_{t\in\mathcal{V}}. \tag{30}$$

Frames with undefined F0 are excluded from $\mathcal{V}$.

SPK-SIM. SPK-SIM measures speaker identity similarity using a pretrained speaker encoder. Let $\phi(\cdot)$ denote the speaker embedding extractor. In our experiments, we use the *resemblyzer* library to compute speaker embeddings $\mathbf{e} = \phi(y)$ and $\hat{\mathbf{e}} = \phi(\hat{y})$, and report cosine similarity:

$$\text{SPK-SIM} = \frac{\mathbf{e}^\top \hat{\mathbf{e}}}{\|\mathbf{e}\|_2 \|\hat{\mathbf{e}}\|_2}. \tag{31}$$

UTMOS. UTMOS is a learned predictor of perceived naturalness and audio quality, trained to approximate human mean-opinion scores (MOS):

$$\mathrm{UTMOS} = g(\hat{y}), \tag{32}$$

where $g(\cdot)$ denotes a pretrained UTMOS model.

Lip Synchronization Metrics. For audio-driven lip synchronization, we evaluate the alignment between the driving audio stream and the generated lip motions using a pretrained audio–visual synchronization network.

Sync-D (Distance). Let $\psi_a(\cdot)$ and $\psi_v(\cdot)$ denote the audio and visual encoders of the sync network, producing embeddings $\mathbf{z}_a$ and $\mathbf{z}_v$. Sync-D is defined as

$$\text{Sync-D} = \|\mathbf{z}_a - \mathbf{z}_v\|_2, \tag{33}$$

where lower values indicate better synchronization.

Sync-C (Confidence). Sync-C reflects the synchronization confidence produced by the sync network:

$$\text{Sync-C} = h(\mathbf{z}_a, \mathbf{z}_v), \tag{34}$$

where $h(\cdot,\cdot)$ denotes the pretrained classifier head. Higher values indicate stronger synchronization confidence.

D Reproducibility and Open Source

To facilitate reproducibility and support further research, we are **progressively releasing** our complete training framework, the model architecture, and the PHV-40 test dataset. We also acknowledge and appreciate the contributions of the open-source community, which have provided essential tools and foundations for this work.

Regarding the evaluation protocols, due to complex environment dependencies among the various third-party assessment tools, we are currently harmonizing these frameworks. The unified evaluation pipeline will be organized and released in the subsequent updates.