A Semantically Consistent Dataset for Data-Efficient Query-Based Universal Sound Separation

Kai Li*, Jintao Cheng*, Chang Zeng, Zijun Yan, Helin Wang, Zixiong Su, Bo Zheng, Xiaolin Hu†

Huggingface

Home Page

1. The Problem: a Data Bottleneck

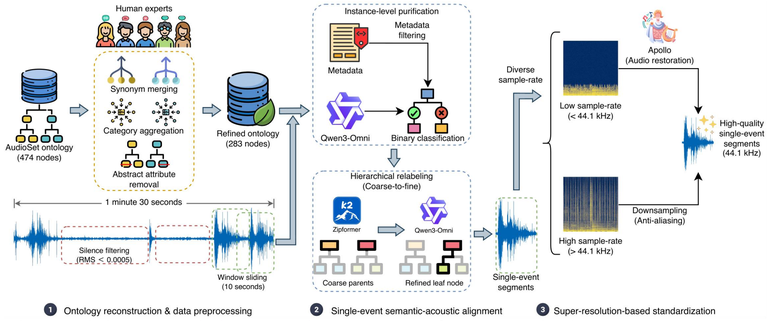
Query-based universal sound separation (USS) isolates a target source from a mixture given a prompt. Yet models still suffer from **residual interference** — background leaks into the output.

Root cause is the data, not the architecture:

- In-the-wild data (AudioSet, VGGSound) have **weak labels** and severe **event co-occurrence**.
- Models learn **spurious correlations** ("rain" ⇒ always wind/traffic) instead of robust acoustic features.
- Brute-force scaling (SAM-Audio ~1M h) is accurate but **prohibitively expensive**.

Can we match SOTA with far less data —
by prioritizing purity over scale?

2. A Single-Event Audio Collection Pipeline



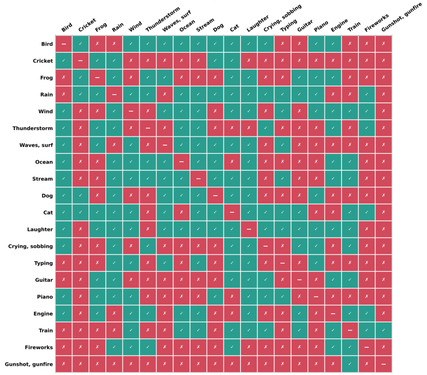
Three coupled stages mine **high-purity single-event** segments from uncurated audio:

- **Ontology reconstruction:** prune AudioSet 474→**283** separable leaf nodes (merge synonyms, drop abstract attributes).
- **Semantic-acoustic alignment:** reject multi-event clips; coarse-to-fine relabel with **Qwen3-Omni** (98% agreement wlth human consensus).
- **Super-resolution standardization:** unify all audio to **44.1 kHz** to enhance signal fidelity and acoustic quality..

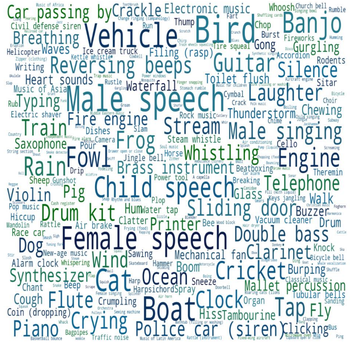
3. The Hive Dataset

Semantically-consistent mixing: A binary co-occurrence matrix *M* from Qwen3-Omni permits only plausible event combinations; tuples are pairwise-compatible.

- **19.6 M** mixtures; cover **2mix-5mix**.
- **292 h** unmixed val+test — broader than prior benchmarks.
- Single-label ✓ Principled mix ✓ Public data & tools ✓



submatrix of the semantic compatibility matrix used in Hive



Overall label distribution visualized as a word cloud

Hive vs. Datasets in Prior USS Research

Research	Train. Dur. (h)	Val. & Test Dur. (h)	Sample-rate	Single-label	Principled mix	Pub. mixed data	Pub. mix tool
LASSNet (Liu et al., 2022)	17.3	~4	32k	✗	✗	✗	✗
CLIPSep (Dong et al., 2023)	550	~50	16k	✗	✗	✗	✗
AudioSep (Liu et al., 2024)	14,100	~109	32k	◐	✗	✗	✓
FlowSep (Yuan et al., 2025)	1,680	~9	16k	◐	✗	✗	✗
OmniSep (Cheng et al., 2025)	560	~10	16k	◐	✗	✗	✗
SAM-Audio (Shi et al., 2025)	~1,000,000	~47	48k	◐	✓	✗	✗
Hive	2,442	292	44.1k	✓	✓	✓	✓

Datasets and mixing pipelines across USS methods.

- **Train. / Val. & Test Dur.**: total duration of the unmixed training set, and of the combined val+test sets.
- **Principled mix**: mixing uses semantic/acoustic constraints, not random combination.
- **Pub. mixed data / mix tool**: pre-mixed data and mixing code are public.

(✓ yes · ◐ partial · ✗ no)

4. Results: Purity Beats Scale

Model (Train Set, Scale)	Signal Fidelity ↑		Perceptual & Semantic Quality				Reference-free Quality ↑			
	SDR	SI-SDR	LPAPS ↓	FAD ↓	CLAP-A ↑	CLAP-T ↑	OQ	Rec	Pre	Fai
MUSDB18-HQ										
AudioSep (Orig., 14.1k h)	-1.01	-1.46	4.94	0.97	0.49	0.09	2.39	3.92	2.40	3.84
FlowSep (Orig., 1.7k h)	–	–	4.72	0.94	0.54	0.11	2.21	3.54	2.30	3.33
SAM-Audio (Orig., 1M h)	–	–	**4.33**	**0.83**	**0.65**	0.12	**3.24**	**4.54**	**3.68**	**4.27**
AudioSep (Hive, 2.4k h)	**1.36**	**0.89**	4.68	0.91	0.58	**0.15**	3.01	4.18	3.14	4.05
FlowSep (Hive, 2.4k h)	–	–	4.34	**0.83**	**0.65**	0.12	3.11	4.22	3.22	3.90
USS-Bench										
AudioSep (Orig., 14.1k h)	-1.86	-3.63	5.02	0.91	0.58	0.11	2.97	4.01	3.04	**4.17**
FlowSep (Orig., 1.7k h)	–	–	5.19	0.93	0.56	0.12	2.33	3.56	2.58	3.14
SAM-Audio (Orig., 1M h)	–	–	4.96	0.90	0.57	0.15	3.47	4.14	**3.77**	3.91
AudioSep (Hive, 2.4k h)	**2.29**	**0.30**	**4.22**	**0.75**	**0.69**	**0.22**	**3.56**	**4.22**	3.68	4.03
FlowSep (Hive, 2.4k h)	–	–	4.62	0.88	0.58	0.15	2.89	3.80	3.34	3.75
VGGClean_eval										
AudioSep (Orig., 14.1k h)	**8.67**	**7.48**	**0.58**	**0.60**	**0.79**	**0.25**	3.42	4.85	3.47	4.52
FlowSep (Orig., 1.7k h)	–	–	1.78	**0.70**	**0.72**	0.14	2.99	4.76	3.03	4.38
SAM-Audio (Orig., 1M h)	–	–	4.27	1.05	0.40	0.21	3.14	3.75	3.72	3.60
AudioSep (Hive, 2.4k h)	7.61	6.43	0.69	0.64	0.71	0.22	**3.44**	**4.91**	**3.48**	**4.54**
FlowSep (Hive, 2.4k h)	–	–	**1.76**	0.81	0.69	**0.18**	**3.18**	**4.85**	**3.79**	**4.47**

Performance comparison on third-party test sets.
"Orig." stands for Original Datasets, and unreported metrics are marked with "–".

(a) AudioSep (SDR ↑).

Model	Co-occ.	Decorr.	Δ
AudioSep (Orig.)	1.65	3.06	−1.41
AudioSep (Hive)	5.48	5.87	−0.39

(b) FlowSep (OQ ↑ and CLAP-T ↑).

Model	OQ			CLAP-T		
	Co-occ.	Decorr.	Δ	Co-occ.	Decorr.	Δ
FlowSep (Orig.)	2.65	2.87	−0.23	0.14	0.18	−0.04
FlowSep (Hive)	3.00	3.09	−0.10	0.18	0.20	−0.02

Controlled Shortcut Analysis

- **Setup:** Paired evaluation isolating statistical co-occurrence bias (Co-occurring vs. Decorrelated backgrounds).
- **Failure Modes of Current USS:** Original baselines rely on shortcuts; SDR drops sharply (Delta = **-1.41 dB**) against unfamiliar backgrounds
- **The Hive Solution:** High-purity supervision eliminates shortcut reliance, shrinking the to just **-0.39 dB**.

5. Takeaways

- Hive-trained models rival **SAM-Audio** using only **~0.2%** of its data, and lead on 3 OOD benchmarks (MUSDB18-HQ, USS-Bench, VGGClean).
- **Purity of supervision > brute-force scale**: a reproducible, low-compute paradigm for robust USS.

arXiv:2601.22599 JusperLee/Hive cslikai.cn/Hive
hf.co/JusperLee/Hive Apache-2.0 · AudioSep/FlowSep ckpts + demo