

USAD 2.0: Scaling Representation Distillation for Universal Audio Understanding

Heng-Jui Chang^{1,*}, Alexander H. Liu¹, Saurabhchand Bhati², Mrudula Athi²,
Anton Ratnarajah², Amit Chhetri², James Glass¹

¹MIT CSAIL, USA ²Amazon, USA

hengjui@mit.edu

Abstract

Audio encoders are critical to modern audio applications as large language models (LLMs) increasingly rely on a single encoder for diverse inputs. While self-supervised learning (SSL) has yielded strong domain-specific encoders like speech or music experts, multi-domain approaches like USAD and SPEAR remain limited in coverage and evaluation. Recent studies also suggest supervised encoders align better with audio LLMs. We present USAD 2.0, a universal encoder integrating knowledge from both SSL and supervised foundation models. USAD 2.0 introduces domain-aware distillation to address teacher mismatch, extends coverage to the music domain, and adds second-stage supervised distillation for downstream use. We further scale the model to one billion parameters via depth scaling. Experiments show USAD 2.0 achieves strong or state-of-the-art performance across probing and LLM-based evaluations.¹

Index Terms: audio representations, self-supervised learning, audio large language models

1. Introduction

Audio encoders have been extensively explored for applications ranging from automatic speech recognition (ASR) to audio codecs [1, 2]. These encoders transform raw waveforms into compact representations, allowing downstream models to access information from audio signals. A widely adopted approach is self-supervised learning (SSL) on large unlabeled datasets, which provides fine-grained features and reduces the reliance on annotated data [3]. However, most SSL models are curated for single-domain usage. E.g., WavLM [4] excels at speech tasks but struggles with out-of-domain audio such as environmental soundscapes. Similar limitations can be observed in general audio [5–9] and music [10–12] SSL models.

With recent advances in audio large language models (LLMs), there is a growing need for strong audio frontends that produce high-quality embeddings across domains, motivating multi-domain audio SSL models. Universal Speech and Audio Distillation (USAD) [13] proposes layer-wise distillation to aggregate knowledge from speech and general-audio SSL encoders. In parallel, Wei et al. distill knowledge from speech and music experts [14], and SPEECH and Audio Representations (SPEAR) distills from multi-codebook vector-quantized SSL models [15]. Nevertheless, these models are primarily evaluated via probing tasks and do not simultaneously cover speech, general audio, and music domains.

Meanwhile, recent studies suggest that *supervised* audio encoders can be particularly effective for audio LLMs, audio re-

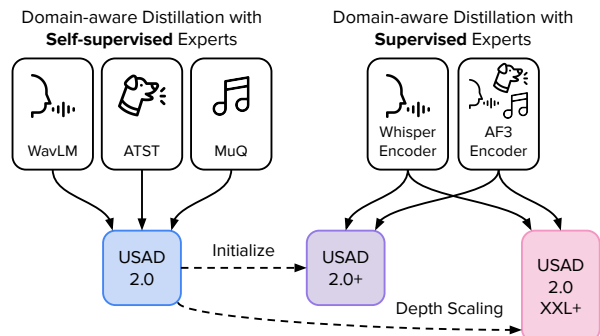


Figure 1: Proposed USAD 2.0. Domain-aware distillation from three SSL experts establishes a strong foundation. Next, supervised experts are distilled to the first stage-initialized encoder. Finally, depth scaling is applied to increase model capacity.

trieval, and speech codecs [16–19]. E.g., the encoder in Audio Flamingo 3 [20] is initialized from Whisper Large [21] and then fine-tuned with joint audio captioning and ASR objectives. With explicit alignment to target applications, such encoders are more likely to succeed as frontends for multimodal LLMs.

In this paper, we build a universal audio encoder that extracts useful representations across multiple audio domains and tasks by distilling from both SSL and supervised audio foundation models. We propose USAD 2.0, which builds on USAD [13] to provide a practical, systematically evaluated framework for integrating domain-specialized audio encoders. As shown in Fig. 1, we first introduce domain-aware distillation, which accounts for whether a teacher matches the input domain. We then incorporate a music teacher along with large-scale music datasets. Next, we propose USAD 2.0+ via second-stage distillation from supervised state-of-the-art teachers to align the encoder with audio LLM applications. Finally, we scale USAD 2.0+ to one billion parameters by reducing temporal resolution and scaling depth with minimal cost. USAD 2.0 achieves superior performance on both probing and LLM-based evaluations across diverse audio domains, demonstrating the effectiveness of the proposed framework as a universal audio encoder. Comprehensive ablation experiments and visualization justify the efficacy of the proposed techniques.

2. Methods

2.1. Recap: USAD

Universal Speech and Audio Distillation (USAD) distills knowledge from two SSL models, one specializing in speech and the other in general audio, into a single encoder for universal audio understanding [13]. USAD uses layer-to-layer

^{**}indicates the corresponding author.

¹<https://hf.co/collections/MIT-SLS/usad2>

knowledge distillation, motivated by the observation that different information types, such as speech content and environmental sounds, are encoded across the hidden layers of SSL models [22]. Although USAD used only two SSL teachers, we generalize the formulation to M teachers. The training loss is the average of per-teacher distillation losses \mathcal{L}_m , each decomposed into layer-wise terms $\mathcal{L}_{m,k}$:

$$\mathcal{L}_{\text{USAD}} = \frac{1}{M} \sum_{m=1}^M \mathcal{L}_m = \frac{1}{MK} \sum_{m=1}^M \sum_{k=1}^K \mathcal{L}_{m,k}, \quad (1)$$

where K is the number of layers from which the student distills for each teacher model. Each $\mathcal{L}_{m,k}$ follows DistilHuBERT [23] by maximizing the similarity between the student and teacher hidden representations. By directly learning the behavior of SSL experts, USAD achieves balanced performance across multiple tasks and data domains. Building on this success, we propose USAD 2.0, which incorporates improved distillation (Sec. 2.2) and scaling techniques (Sec. 2.3) to support a broader range of audio understanding applications.

2.2. USAD 2.0

2.2.1. Domain-aware Distillation

This section introduces domain-aware distillation to improve USAD. Each SSL teacher specializes in a specific audio domain, but USAD weights all teacher losses equally, regardless of the input. To encourage higher-quality representations, we upweight the loss when the input domain matches the corresponding teacher domain. Assuming the M teachers each specialize in a unique domain, the loss for an instance from domain m_{data} is

$$\mathcal{L}_{\text{USAD 2.0}} = \sum_{m=1}^M w_m(m_{\text{data}}) \mathcal{L}_m, \quad (2)$$

where $w_m(m_{\text{data}})$ scales the contribution of the m^{th} teacher. We introduce a scaling factor $\alpha > 1$ to control the ratio between matched and mismatched domains. Enforcing $\sum_{m=1}^M w_m(m_{\text{data}}) = 1$, we define

$$w_m(m_{\text{data}}) = \begin{cases} \frac{\alpha}{\alpha + M - 1}, & m = m_{\text{data}} \\ \frac{1}{\alpha + M - 1}, & m \neq m_{\text{data}} \end{cases} \quad (3)$$

When $\alpha = 1$, the weights reduce to $\frac{1}{M}$ for all teachers. If m_{data} is unknown, we also set $w_m = \frac{1}{M}$. Unlike [14], which effectively takes $\alpha \rightarrow \infty$, our *soft* weighting still allows distillation from mismatched teachers. This is beneficial when domains share structure. E.g., since speech often appears in mixed audio, distilling from a speech teacher can help the student acquire denoising capability. Thus, mismatched teachers remain active with smaller weights, allowing the student to retain cross-domain cues while still emphasizing the most relevant expert for each input domain.

2.2.2. Music Domain Expert

Empirically, USAD underperforms music SSL models on music-centric tasks such as genre and key classification, likely due to the lack of music-domain supervision. Given the growing importance of music-focused SSL methods and applications [10–12,24], we introduce a music-domain expert and additional music audio data for USAD 2.0 to distill from. Combined with domain-aware distillation, this gives USAD 2.0 a broader and more diverse skill set.

2.3. Scaling USAD 2.0

2.3.1. Second-stage Distillation with Supervised Experts

Recent progress in audio LLMs has highlighted the effectiveness of audio encoders pre-trained with supervised objectives like ASR. In particular, many audio encoders are fine-tuned from Whisper’s encoder [16, 20, 21, 24]. Hence, we propose USAD 2.0+ via second-stage distillation from state-of-the-art supervised audio encoders. We first identify the strongest experts using probing and LLM-based evaluations: the Whisper Large encoder for multilingual speech [21] and the Audio Flamingo 3 encoder for general audio understanding [20]. We then initialize USAD 2.0+ from the SSL-distilled student and distill from the final layers of both supervised teachers. This stage aligns USAD 2.0 with audio LLMs while preserving the fine-grained representations characteristic of SSL pre-training.

2.3.2. Efficient Model Size Scaling

Scaling the audio encoder can improve downstream performance by increasing model capacity, but incurs substantially higher computational cost. We therefore propose two simple approaches to scale USAD 2.0: temporal resolution reduction and depth scaling. First, since the sequence length processed by self-attention dominates training and inference cost, we reduce the feature framerate from 50Hz to 25Hz with a $2 \times$ CNN feature extractor stride. Although the temporal resolution is reduced, increasing the number of layers and hidden dimensions can still improve the encoder’s overall capacity and capability. Second, we reuse the weights of a pre-trained USAD 2.0 model, apply depth scaling, and train the expanded model for only a few more updates. Specifically, we scale our XLarge model from 32 to 48 layers with depth up-scaling [25] by copying and stacking the first and last 24 layers. These methods avoid training large models from scratch and enable scaling USAD 2.0 to 1B parameters within an academic budget.

3. Experiments

3.1. Setup

3.1.1. USAD 2.0 Training

Following USAD [13], we create a multi-domain audio dataset by combining various multilingual speech (116K hours) [28–42], general audio (21K hours) [43–46], and music (13K hours) [47–52] corpora. The domain labels are assigned according to each dataset’s original purpose, and the domain-aware distillation scale α is set to 10 for all models. USAD 2.0 follows the same architecture as USAD [13], except that the XLarge and XXL models use a 25Hz framerate for efficiency. For first-stage training, USAD 2.0 distills from WavLM [4], ATST-Frame [8], and MuQ [12], respectively representing speech, audio, and music experts. Thus, this stage directly evaluates the extension from the two-teacher setting of USAD to three SSL teachers. Supervised distillation uses the Whisper Large-v3 encoder [21] and Audio Flamingo 3 (AF3) AF-Whisper [20] as targets, distilling only the last layer of each expert due to their supervised nature. Because AF3 is multi-domain, the losses from all domains are treated equally. The first and second stages are trained with 600K and 50K updates, respectively.

3.1.2. Evaluation

We include multiple protocols to evaluate the proposed models, ranging from simple probing tasks to audio LLM evalua-

Table 1: Results on HEAR [26], MARBLE [27], and XARES-LLM. All reported numbers are obtained by using only the audio encoder of each model. E.g., the decoder of each Whisper model is discarded. The best results are shown in **bold**, and the second- and third-best results are underlined.

Encoder	Params	HEAR	MARBLE	XARES-LLM	
		Avg	Avg	Track A	Track B
Single-encoder SOTA					
Base	~90M	80.6	74.0	0.660	0.418
		<u>SPEAR</u>	<u>MERT</u>	<u>SPEAR</u>	<u>Whisper</u>
		Base [15]	95M [10]	Base [15]	Small [21]
Large	~300M	81.8	77.0	0.691	0.454
		<u>SPEAR</u>	<u>MuQ</u> [12]	<u>SPEAR</u>	<u>Whisper</u>
		Large [15]		Large [15]	Medium [21]
XLarge	~600M	82.6	75.1	<u>0.782</u>	0.457
		<u>SPEAR</u>	<u>SPEAR</u>	<u>AF3</u> [20]	<u>Whisper</u>
		XLarge [15]	XLarge [15]		Large [21]
Multi-expert Encoder (USAD 2.0 Teachers)					
Self-supervised	734M	82.0	<u>76.1</u>	0.645	0.462
(WavLM + ATST + MuQ)					
Supervised	1274M	81.8	72.4	0.806	0.685
(Whisper + AF3)					
USAD 2.0 (Self-supervised Teachers)					
Small	25M	81.0	72.9	0.604	0.357
Base	97M	81.9	74.1	0.645	0.442
Large	336M	82.9	<u>75.8</u>	0.667	0.473
XLarge	695M	82.5	75.7	0.708	0.485
USAD 2.0+ (Supervised Teachers)					
Large+	336M	<u>84.0</u>	75.1	0.769	0.580
XLarge+	695M	84.4	75.0	0.772	<u>0.611</u>
XXLarge+	1036M	84.4	75.6	<u>0.783</u>	<u>0.624</u>

tions. **HEAR** is a benchmark that probes frozen SSL model representations for various tasks, covering speech, sound, and music [26]. **MARBLE** is a music-focused probing benchmark similar to HEAR [27]. Finally, we follow **XARES-LLM** (The Interspeech 2026 Audio Encoder Capability Challenge for Large Audio Language Models) by training a multitask audio LLM using frozen representations of audio encoders [53]. Track A (classification tasks) covers keyword spotting, speaker/language identification, spoof detection, intent/emotion/sound/genre/instrument classification, and sound event detection. Track B (understanding tasks) includes English/Mandarin ASR and audio/music captioning. To ensure controlled comparison of audio representations, all baselines are evaluated in an encoder-only setting, including prior universal or multi-domain encoders [13, 15], domain-specialized SSL models [4, 8, 12], supervised audio LLM-oriented encoders [20, 21, 24], and multi-expert teacher topline.

3.2. Main Results

Tab. 1 reports average scores of each benchmark. On HEAR, the unsupervised USAD 2.0 models consistently outperform prior state-of-the-art models of comparable sizes. Although reducing the framerate to 25Hz slightly degrades the XLarge model performance, the score remains competitive with SPEAR XLarge [15]. Introducing the second-stage distillation with supervised teachers (USAD 2.0+) yields further improvements, pushing beyond the prior state-of-the-art model.

For the music-centric evaluation on MARBLE, USAD 2.0 demonstrates robust multi-domain coverage. The unsupervised Large model surpasses both the Base and XLarge baselines, while still highly competitive with specialized, music-

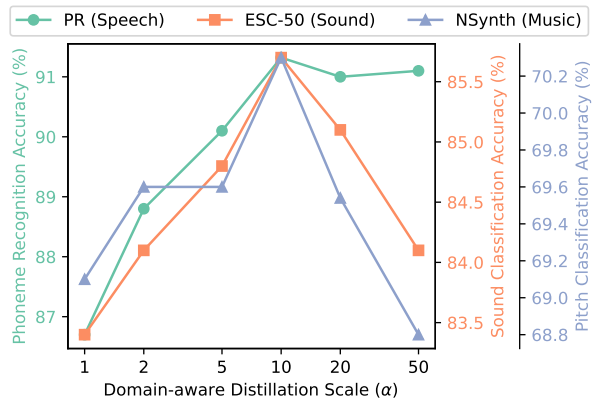


Figure 2: Domain-aware distillation scale vs. phoneme recognition, sound classification, and pitch classification, where $\alpha = 10$ is most robust across domains.

Table 2: Ablation studies on phoneme recognition (PR) [28, 54], sound classification (ESC-50) [55], and pitch classification (NSynth) [50]. All models use the Small 25M-parameter backbone without fine-tuning.

Method	PR PER↓	ESC-50 Acc↑	NSynth Acc↑
USAD [13]	8.8	80.3	55.1
USAD 2.0 (proposed)	8.7	85.7	70.3
w/o Domain-aware Distillation	13.3	83.4	69.1
w/o Music Domain Teacher	8.5	85.2	49.1
w/o Music Data	8.4	84.3	53.2

only models like MuQ [12]. The supervised variants maintain this strong performance, indicating that aligning with supervised experts preserves fine-grained music understanding.

On XARES-LLM, USAD 2.0 exhibits highly effective scaling. The supervised USAD 2.0+ variants match or surpass the top-performing XLarge single-encoder baselines, especially on Track B. The most significant gains are observed in Track B (understanding). The results indicate the usefulness of distilling from supervised experts.

To provide a performance topline, we include the “Multi-expert Encoder” results, which are obtained by concatenating the outputs of the teacher models. While these ensembles achieve high scores, they require significantly more parameters than our distilled students. Notably, USAD 2.0 models often match or exceed these multi-expert teachers while maintaining a much smaller parameter footprint. By integrating the fine-grained acoustic details of SSL experts with the high-level semantic alignment of supervised models, USAD 2.0 establishes a new state-of-the-art for efficient, highly capable universal audio encoders across speech, sound, and music.

3.3. Ablation Studies

This section ablates and analyzes the proposed techniques for USAD 2.0 training and scaling. As shown in Fig. 2, the domain-aware distillation scale α achieves the best performance across domain tasks when set to 10. When α is too small, the student learns from weaker targets due to mismatches between the expert and the data domain in the USAD approach [13]. Meanwhile, an overly large α degrades performance, indicating that excessively strong supervision from matched-domain experts can reduce the cross-domain generalizability.

Table 3: Ablation studies on the initialization approaches.

USAD 2.0+	XARES-LLM	
	Track A	Track B
XLarge+		
from scratch	0.731	0.574
init with XLarge (proposed)	0.772	0.611
XXLarge+ (init with depth-scaled XLarge)		
new top 16 layers	0.779	0.617
uniform layer duplication	0.775	0.622
depth up-scaling [25]	0.783	0.624

Table 4: Inference efficiency of different model sizes. The metrics are measured on an A5000 GPU and averaged over 50 runs, with a 30-second audio input.

USAD 2.0 Size	Params	Framerate	RTF↓	Peak GPU Memory↓
Large	336M	50Hz	0.0029	1.2GB
XLarge (50Hz)	695M	50Hz	0.0051	2.2GB
XLarge (25Hz, proposed)	695M	25Hz	0.0018	1.7GB
XXLarge (50Hz)	1036M	50Hz	0.0077	3.0GB
XXLarge (25Hz, proposed)	1036M	25Hz	0.0026	2.4GB

We conduct ablation studies on the proposed methods in Tab. 2. USAD 2.0 surpasses USAD [13] under the same training and data setup. Next, results without domain-aware distillation further demonstrate the importance of this technique for balancing performance across domains. Without the music-domain teacher for distillation, the pitch classification accuracy drops by 30% (relative). A similar phenomenon is observed when music data is removed, implying the necessity of both domain experts and in-domain data for music.

Furthermore, we evaluate the proposed initialization and depth-scaling approaches in Tab. 3. First, initializing the supervised XLarge+ model from the SSL-pretrained XLarge backbone yields substantial gains on the XARES-LLM benchmark compared to training from scratch, improving Track A from 0.731 to 0.772 and Track B from 0.574 to 0.611. Next, we investigate three methods for scaling the 32-layer model to a 48-layer XXLarge+ architecture. Specifically, “new top 16 layers” appends randomly initialized transformer encoder layers on top of the original model; “uniform layer duplication” duplicates every even-numbered layer for a 1.5× expansion; and “depth up-scaling” follows [25] by copying and stacking the first 24 and last 24 layers. These depth-scaled models are trained via domain-aware distillation from supervised teachers. The increased capacity allows all depth-scaling variants to outperform the XLarge+ baseline, with depth up-scaling achieving the highest overall performance on both Tracks A and B. Collectively, these ablation studies confirm the efficacy of the proposed USAD 2.0 training and scaling strategies.

3.4. Inference Efficiency

We compare the inference efficiency of USAD 2.0 to assess its real-world applicability. We use the real-time factor (RTF), defined as the ratio of inference time to input audio duration, to quantify inference speed. As shown in Tab. 4, the Large model has the lowest memory usage because of the model size. Furthermore, the proposed framerate reduction yields substantial computational benefits. Specifically, the 25Hz XLarge model speeds up by more than 2.8× and reduces memory usage by

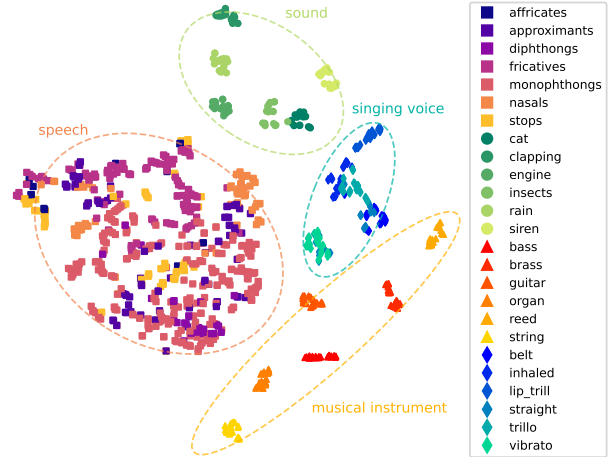


Figure 3: *t*-SNE [56] visualization of USAD 2.0 XXLARGE+ hidden representations with speech [57], environmental sounds [55], musical instruments [50], and singing voices [58].

20% compared with the 50Hz counterpart. The 25Hz XXLarge model, despite scaling to over one billion parameters, operates faster than the 336M-parameter Large model at 50Hz, while keeping peak memory usage at a manageable 2.4GB. Taken together with Tab. 1 and Tab. 4, these results show that the efficient scaling of USAD 2.0 yields continual improvements while maintaining fast inference performance.

3.5. Representation Visualization

This section visualizes the hidden representations of USAD 2.0 XXLARGE+ to understand how audio is encoded into high-dimensional embedding spaces. We visualize the 40th-layer embeddings of USAD 2.0 XXLARGE+, which achieves the best XARES-LLM performance. The embeddings are mean-pooled along the temporal dimension for each audio clip, except for speech, where each phoneme segment is pooled.

As shown in Fig. 3, the embeddings form four distinct macro-clusters corresponding to the broad input domains: speech, environmental sound, singing voice, and musical instrument. Within each domain, representations are further organized into fine-grained categories. In particular, environmental sounds and musical instruments form tightly isolated sub-clusters, while speech phonemes exhibit a more continuous distribution with slight overlap, reflecting the connected nature of spoken articulation. Moreover, singing voice representations also cluster by vocal techniques with some overlap. These observations indicate the model effectively disentangles multiple input domains while preserving intra-domain categorical structure, offering well-separated representations that allow downstream models easy access to the required information.

4. Conclusion

This paper presents USAD 2.0, a scalable universal audio encoder for audio LLMs, combining domain-aware distillation, a music expert, and supervised distillation to integrate strengths from self-supervised and supervised foundation models. Efficient approaches scale the model to 1B parameters within an academic budget while maintaining fast inference. USAD 2.0 delivers strong, balanced cross-domain performance, outperforming prior universal and domain-specific encoders, making it a practical frontend for next-generation audio LLMs.

5. Generative AI Use Disclosure

Generative AI is used to polish the manuscript without significant changes to the authors' original draft.

6. References

- [1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.
- [2] H.-J. Chang, H. Gong, C. Wang, J. Glass, and Y.-A. Chung, "Dc-spin: A speaker-invariant speech tokenizer for spoken language models," in *Interspeech*, 2025.
- [3] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang *et al.*, "A large-scale evaluation of speech foundation models," *TASLP*, 2024.
- [4] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE JSTSP*, vol. 16, 2022.
- [5] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *ICML*, 2022.
- [6] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "Eat: Self-supervised pre-training with efficient audio transformer," in *IJ-CAI*, 2024.
- [7] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech*, 2024.
- [8] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *TASLP*, 2024.
- [9] T. Alex, S. Atito, A. Mustafa, M. Awais, and P. J. Jackson, "Sslam: Enhancing self-supervised models with audio mixtures for polyphonic soundscapes," in *ICLR*, 2025.
- [10] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *ICLR*, 2024.
- [11] M. Won, Y.-N. Hung, and D. Le, "A foundation model for music informatics," in *ICASSP*, 2024.
- [12] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, "Muq: Self-supervised music representation learning with mel residual vector quantization," *arXiv*, 2025.
- [13] H.-J. Chang, S. Bhati, J. Glass, and A. H. Liu, "USAD: Universal speech and audio representation via distillation," in *ASRU*, 2025.
- [14] J.-C. Wei, Y.-C. Lin, F. Ritter-Gutierrez, and H.-y. Lee, "Multi-distillation from speech and music representation models," in *ASRU*, 2025.
- [15] X. Yang, Y. Yang, Z. Jin, Z. Cui, W. Wu, B. Li, C. Zhang, and P. Woodland, "Spear: A unified ssl framework for learning speech and audio representations," *arXiv*, 2025.
- [16] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv*, 2024.
- [17] H. Dinkel, G. Li, J. Liu, J. Luan, Y. Niu, X. Sun, T. Wang, Q. Xiao, J. Zhang, and J. Zhou, "Midashenglm: Efficient audio understanding with general audio captions," *arXiv*, 2025.
- [18] Y. Song, L. Zhang, C. Wu, A. Liu, W. Jia, H. Wang, and X. Zhou, "Stabletoken: A noise-robust semantic speech tokenizer for resilient speechllms," *arXiv*, 2025.
- [19] A. Vyas, H.-J. Chang, C.-F. Yang, P.-Y. Huang, L. Gao, J. Richter, S. Chen, M. Le, P. Dollár, C. Feichtenhofer *et al.*, "Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning," in *CVPR*, 2026.
- [20] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv*, 2025.
- [21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *ICML*, 2023.
- [22] H.-J. Chang, N. Dong, R. Mavlyutov, S. Popuri, and Y.-A. Chung, "CoLLD: Contrastive layer-to-layer distillation for compressing multilingual pre-trained speech encoders," in *ICASSP*, 2024.
- [23] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "DistilHuBERT: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP*, 2022.
- [24] S. Ghosh, A. Goel, L. Koroshinadze, S.-g. Lee, Z. Kong, J. F. Santos, R. Duraiswami, D. Manocha, W. Ping, M. Shoenybi *et al.*, "Music flamingo: Scaling music understanding in audio language models," *arXiv*, 2025.
- [25] S. Kim, D. Kim, C. Park, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim *et al.*, "Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling," in *NAACL*, 2024.
- [26] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*, 2022.
- [27] R. Yuan, Y. Ma, Y. Li, G. Zhang, X. Chen, H. Yin, Y. Liu, J. Huang, Z. Tian, B. Deng *et al.*, "Marble: Music audio representation benchmark for universal evaluation," *NeurIPS*, 2023.
- [28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.
- [29] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020.
- [30] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech*, 2020.
- [31] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, 2020.
- [32] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *ACL*, 2021.
- [33] W. Chen, W. Zhang, Y. Peng, X. Li, J. Tian, J. Shi, X. Chang, S. Maiti, K. Livescu, and S. Watanabe, "Towards robust speech representation learning for thousands of languages," in *EMNLP*, 2024.
- [34] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *JMLR*, vol. 25, no. 97, 2024.
- [35] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," in *Interspeech*, 2021.
- [36] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *SLT*, 2021.
- [37] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text," in *LREC*, 2004.
- [38] P. Roach, S. Arnfield, W. Barry, S. Dimitrova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, A. M. Hallum, A. Lamel *et al.*, "Babel: a database of central and eastern european languages," in *LREC*, 1998.
- [39] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Babna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *SLT*, 2023.
- [40] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *O-COCOSDA*, 2017.

- [41] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*, 2015.
- [42] T. A. Nguyen, W.-N. Hsu, A. D'Avirro, B. Shi, I. Gat, M. Fazel-Zarani, T. Remez, J. Copet, G. Synnaeve, M. Hassid, F. Kreuk, Y. Adi, and E. Dupoux, "Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis," in *Interspeech*, 2023.
- [43] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NIPS*, 2016.
- [44] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.
- [45] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023.
- [46] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.
- [47] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," in *ISMIR*, 2016.
- [48] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The mtg-jamendo dataset for automatic music tagging," in *MLAMD at ICML*, 2019.
- [49] I. A. P. Santana, F. Pinhelli, J. Donini, L. Catharin, R. B. Mangolin, V. D. Feltrim, M. A. Domingues *et al.*, "Music4all: A new music database and its applications," in *IWSSIP*, 2020.
- [50] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *ICML*, 2017.
- [51] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *ISMIR*, 2009.
- [52] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the maestro dataset," in *ICLR*, 2019.
- [53] H. Dinkel, J. Zhou, G. Wang, Y. Niu, J. Zhang, Y. Hao, Y. Liu, K. Li, W. Wang, Z. Wu *et al.*, "The interspeech 2026 audio encoder capability challenge for large audio language models," *arXiv*, 2026.
- [54] S.-w. Yang *et al.*, "SUPERB: Speech processing universal performance benchmark," in *Interspeech*, 2021.
- [55] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *ACM MM*, 2015.
- [56] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *JMLR*, 2008.
- [57] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *LDC*, 1993.
- [58] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *ISMIR*, 2018.

A. Training Setup

A.1. Data

As shown in Tab. 5, we construct a large multi-domain audio dataset by combining publicly available multilingual speech, general audio, and music corpora. The first-stage SSL distillation uses all datasets in the first three sections of Tab. 5 to encourage domain diversity. For second-stage supervised distillation, we remove several smaller and noisier datasets to stabilize training. Since fine-grained speech representations are more difficult to distill than sound and music representations, speech accounts for roughly half of the training data. Because the model already acquires strong music capability during the first stage, we reduce the music proportion from 15% to 9% in the second stage.

A.2. USAD 2.0 Training

The complete hyperparameters are reported in Tab. 6. Following USAD [13], the audio waveform is first converted into 128-bin Mel spectrogram features. The USAD 2.0 backbone comprises a two-layer CNN feature extractor, a five-layer convolutional positional encoding module [1], and a transformer encoder. For first-stage training, USAD 2.0 distills from WavLM [4], ATST-Frame [8], and MuQ [12], which serve as domain experts for speech, general audio, and music, respectively. This stage directly evaluates the extension from the two-teacher setting of USAD to three SSL teachers. For second-stage supervised distillation, the targets are the Whisper Large-v3 encoder [21] and Audio Flamingo 3 (AF3) AF-Whisper [20]. Only the last layer of each supervised expert is distilled, and losses from all domains are weighted equally because AF3 is multi-domain. The domain-aware distillation scale α is set to 10 for all models.

B. Additional Results

B.1. Probing and Fine-tuning Benchmarks

We provide complete experimental results for these tasks:

- Audio Tagging and Sound Classification: Tab. 7.
- HEAR [26]: Tab. 8.
- MARBLE [27]: Tab. 9.
- SUPERB [54]: Tab. 10.

B.2. XARES-LLM Benchmark

Following the XARES-LLM benchmark [53], we evaluate several state-of-the-art audio encoders across different domains and report in Fig. 4. Excluding multi-expert encoders, USAD 2.0 XXLarge+ is the best-performing encoder on both Track A and Track B. Scaling from 0.3B (Large+) to 1B (XXLarge+) parameters yields consistent improvements across most tasks. Comparing SSL-based and supervised encoders shows that supervised encoders are generally stronger for audio LLM applications, supporting the motivation for second-stage supervised distillation in Sec. 2.3.1. Moreover, domain-specific encoders show strong in-domain capabilities but weaker out-of-domain performance. For example, WavLM Large [4] performs well on several speech-related tasks, whereas the Music Flamingo encoder [24] largely loses speech processing ability, especially for ASR (AISHELL-1 and LibriSpeech), after training with more music data. In contrast, USAD 2.0 XXLarge+ exhibits more balanced performance across domains and tasks. Overall, these

Table 5: *Datasets for USAD 2.0 training. The dataset sizes might differ from the original ones due to preprocessing. \blacklozenge indicates the datasets removed after the first-stage training.*

Dataset	Clips	Duration (hours)	Proportion
Speech			
LibriVox [28–30]	13,841,487	55,608	42.24%
Common Voice 17 [31]	5,455,997	9,023	16.65%
VoxPopuli (English) [32]	3,051,826	24,084	9.31%
MMS unlabeled v2 [33, 34] \blacklozenge	2,904,325	7,410	8.86%
GigaSpeech [35]	2,568,818	5,306	7.84%
VoxLingua107 [36]	2,076,677	5,726	6.34%
MLS (non-English) [30]	1,445,342	6,027	4.41%
Fisher [37] \blacklozenge	623,915	1,277	1.90%
BABEL [38]	381,220	798	1.16%
FLEURS [39]	267,042	951	0.81%
AISHELL-1 [40]	69,958	105	0.21%
CHiME-3 [41] \blacklozenge	50,180	111	0.15%
Zereth-Korean \blacklozenge	17,892	41	0.05%
Expresso [42] \blacklozenge	14,468	31	0.04%
General Audio			
SoundNet [43] \blacklozenge	4,637,914	12,385	57.52%
AudioSet [44]	1,929,338	5,318	23.93%
LAION-Audio-630k [45]	1,311,739	3,438	16.27%
VGGSound [46]	183,721	510	2.28%
Music			
FMA [47]	2,961,775	8,173	58.14%
MTG-Jamendo [48]	1,365,876	3,766	26.81%
Music4All [49]	327,807	911	6.44%
NSynth [50]	289,205	321	5.68%
MagnaTagATune [51]	77,580	209	1.52%
MAESTRO [52]	71,847	199	1.41%
Multi-domain Dataset (Self-supervised Teachers)			
Speech (1 \times upsample)	32,769,147	116,495	48.80%
General Audio (3 \times upsample)	24,188,136	64,954	36.02%
Music (2 \times upsample)	10,188,180	27,158	15.17%
Total	67,145,463	208,607	100.00%
Multi-domain Dataset (Supervised Teachers)			
Speech (1 \times upsample)	29,158,367	107,627	53.21%
General Audio (6 \times upsample)	20,548,788	55,598	37.50%
Music (1 \times upsample)	5,094,090	13,579	9.30%
Total	54,801,245	176,805	100.00%

results highlight the effectiveness of the proposed distillation framework and the usefulness of USAD 2.0 as a universal encoder for audio LLM applications.

B.3. Representation Visualization

As shown in Fig. 5, we provide additional hidden-layer visualizations of USAD 2.0 XXLarge+, complementing Fig. 3. Speech, sound, and music embeddings are already separated into several macro-clusters in the lower layers. The main difference between lower and upper layers is observed in the speech representations (shown as squares): lower layers tend to keep phonemes within the same category closer together, whereas upper layers mix different phoneme categories more heavily. This suggests that the lower layers retain behavior similar to speech SSL models [4], likely due to first-stage SSL distillation, while the upper layers are further aligned with supervised experts through second-stage distillation.

Table 6: Hyperparameters of USAD 2.0.

	Small Ablation	Small	Base	Large	XLarge	Large+	XLarge+	XXLarge+
Model								
Initialization	–	–	–	–	–	USAD 2.0 Large	USAD 2.0 XLarge	USAD 2.0 XLarge
Parameters	25M	25M	97M	336M	695M	336M	695M	1036M
Hidden Size	384	384	768	1024	1280	1024	1280	1280
FFN Size	1536	1536	3072	4096	5120	4096	5120	5120
Layers	12	12	12	24	32	24	32	48
Heads	6	6	12	16	20	16	20	20
Framerate	50Hz	50Hz	50Hz	50Hz	25Hz	50Hz	25Hz	25Hz
Teachers								
Speech	WavLM Base+	WavLM Base+	WavLM Base+	WavLM Large	WavLM Large	Whisper Large-v3	Whisper Large-v3	Whisper Large-v3
Audio	ATST Frame	ATST Frame	ATST Frame	ATST Frame	ATST Frame	Audio Flamingo 3	Audio Flamingo 3	Audio Flamingo 3
Music	MuQ _{iter}	MuQ _{iter}	MuQ _{iter}	MuQ _{iter}	MuQ _{iter}	–	–	–
Training								
Optimizer	Adam							
Learning Rate	5e-4	8e-4	1.2e-3	1.5e-3	2e-3	5e-4	5e-4	5e-4
LR Warmup	4k	32k	32k	32k	32k	4k	4k	4k
Updates	150k	600k	600k	600k	600k	50k	50k	50k
Batch Size	200s	800s	800s	1200s	1200s	800s	800s	800s
A6000 GPUs	1	4	4	4	4	4	4	4

Table 7: Results on audio tagging (AS-20K) [44] and sound classification (ESC-50) [55]. The audio encoders are fully fine-tuned.

Model	Params	AS-20K mAP \uparrow	ESC-50 Acc \uparrow
Audio SSL Models			
BEATs iter3 [5]	90M	38.3	95.6
EAT [6]	88M	40.2	95.9
SSLAM [9]	88M	40.9	96.2
ATST Frame [8]	86M	39.0	91.1
Multi-domain Audio Models			
USAD Small [13]	25M	34.5	89.3
USAD Base [13]	97M	35.7	91.1
USAD Large [13]	336M	37.4	92.7
SPEAR Base [15]	94M	39.1	–
SPEAR Large [15]	327M	39.2	–
SPEAR XLarge [15]	600M	39.4	–
USAD 2.0 (Self-supervised Teachers)			
USAD 2.0 Small	25M	37.2	93.5
USAD 2.0 Base	97M	40.0	94.6
USAD 2.0 Large	336M	40.1	95.0
USAD 2.0+ (Supervised Teachers)			
USAD 2.0 Large+	336M	40.6	96.8
USAD 2.0 XLarge+	695M	40.9	96.4

Table 8: Results on HEAR [26].

Model	Params	BJ	CD	D16	E50	F50k	Gun	GZ-Gen	GZ-M/S	LiCt	MST	Mri-S	Mri-T	SPC-5	SPC-F	NS-5	NS-50	VI	VL	Average			
																				Speech	Env	Music	All
Audio Models																							
ATST Frame [8]	86M	95.8	76.7	95.7	89.0	55.7	94.3	88.3	100.0	78.1	24.4	97.5	96.9	92.6	95.1	68.6	82.0	22.3	66.9	72.0	83.7	81.7	78.9
Dasheng Base [7]	86M	93.6	78.7	95.7	82.9	51.0	92.9	89.2	99.2	76.6	43.9	96.1	94.9	95.9	97.1	71.8	83.3	16.7	69.9	72.5	80.6	84.0	79.4
Dasheng 0.6B [7]	630M	94.9	81.2	94.4	85.9	53.9	97.6	88.6	97.6	80.7	43.5	96.6	96.2	97.0	97.5	74.6	85.8	17.8	74.7	74.8	83.0	84.7	81.0
Dasheng 1.2B [7]	1.2B	96.2	81.6	94.2	85.3	54.2	99.1	88.8	99.1	79.6	43.3	96.8	96.1	97.1	97.9	74.4	85.6	19.4	78.7	75.7	83.2	84.9	81.4
Multi-domain Audio Models																							
USAD Small [13]	25M	94.5	78.2	89.5	81.8	51.1	93.2	86.6	98.5	77.0	25.3	97.3	94.3	96.2	97.2	55.6	77.7	20.0	73.6	73.7	78.9	78.7	77.1
USAD Base [13]	97M	95.8	80.0	93.6	82.2	52.2	94.0	86.3	100.0	78.7	26.7	97.3	95.7	96.6	97.6	57.0	81.6	19.5	76.0	74.7	80.5	80.0	78.4
USAD Large [13]	336M	94.1	79.5	93.9	83.4	53.0	97.6	87.4	100.0	79.1	38.4	97.4	96.1	97.0	97.5	57.0	83.2	18.5	75.3	74.5	82.0	81.7	79.4
SPEAR Base [15]	94M	95.3	82.0	95.1	85.9	54.2	95.2	88.8	100.0	76.2	26.8	97.2	96.0	97.3	98.2	69.4	82.2	24.6	85.6	77.3	82.6	82.0	80.6
SPEAR Large [15]	327M	94.9	83.8	95.9	87.6	56.4	97.6	89.2	99.2	78.7	27.9	97.4	97.5	98.1	98.3	70.2	85.3	25.7	88.5	78.9	84.4	82.7	81.8
SPEAR XLarge [15]	600M	95.3	83.6	96.0	89.4	57.1	96.3	91.0	100.0	80.7	27.7	97.4	97.9	98.4	98.6	74.2	86.0	26.6	90.4	79.7	84.7	83.7	82.6
Whisper Large-v3 [21]	637M	95.3	81.8	93.4	86.1	42.4	92.0	87.6	96.0	70.6	24.5	97.4	94.9	98.5	98.6	62.4	66.7	25.6	98.5	78.9	78.5	78.1	78.5
AF3 Whisper [20]	637M	96.2	85.5	93.8	94.3	61.4	97.6	94.8	96.9	71.8	28.7	97.8	97.1	97.4	98.2	71.4	81.8	25.7	92.9	78.6	86.8	83.1	82.4
Multi-expert Encoder																							
Self-supervised																							
(WavLM + ATST + MuQ)	734M	95.8	80.3	93.6	83.4	54.5	97.6	86.8	97.7	79.7	42.4	97.7	98.1	96.6	97.7	75.8	88.7	24.6	84.1	77.2	82.3	85.4	82.0
Supervised																							
(Whisper + AF3)	1274M	95.3	85.2	93.4	92.3	54.7	95.2	92.8	96.1	70.6	27.9	97.8	96.1	98.1	98.4	71.4	81.8	27.1	98.4	79.6	83.9	82.4	81.8
USAD 2.0																							
USAD 2.0 Small	25M	95.8	77.8	94.3	85.3	53.7	92.9	87.3	100.0	78.2	46.4	97.3	97.5	96.3	97.0	76.8	87.4	21.5	72.3	73.9	81.5	86.1	81.0
USAD 2.0 Base	97M	95.3	78.7	95.1	86.6	55.7	92.3	89.9	100.0	80.1	49.4	97.7	97.9	96.1	96.7	79.2	89.3	22.2	72.2	74.3	82.4	87.3	81.9
USAD 2.0 Large	336M	96.2	79.1	95.0	87.7	56.0	97.6	91.1	100.0	80.6	50.5	97.9	98.3	96.4	97.1	80.6	89.9	24.1	73.6	75.1	84.1	88.1	82.9
USAD 2.0 XLarge	695M	94.9	79.9	96.2	88.8	57.1	95.2	89.5	100.0	79.4	48.9	97.4	98.4	95.6	97.1	80.4	90.4	24.4	71.6	74.7	84.4	87.5	82.5
USAD 2.0+																							
USAD 2.0 Large+	336M	96.2	80.8	95.3	93.1	62.2	96.7	91.4	100.0	79.6	43.4	97.9	97.6	97.8	98.1	73.6	87.7	26.1	95.0	79.5	86.8	86.0	84.0
USAD 2.0 XLarge+	695M	96.6	81.9	95.3	93.8	62.8	95.2	91.9	100.0	79.4	43.4	97.7	98.0	97.8	98.0	76.4	88.2	26.2	96.0	79.9	86.8	86.5	84.4
USAD 2.0 XXLarge+	1036M	96.2	82.6	95.0	93.8	62.5	96.4	93.2	100.0	79.7	42.8	97.5	97.9	97.5	97.8	75.6	88.1	25.8	96.3	79.9	86.9	86.4	84.4

Table 9: Results on MARBLE [27].

Model	Params	GTZAN Genre	GS Key	EMO Emotion		VocalSet Singer Tech		MTT Tagging		NSynth Instr. Pitch		
				Acc \uparrow	Acc \uparrow	R2 \uparrow	R2 \uparrow	Acc \uparrow	Acc \uparrow	ROC \uparrow	AP \uparrow	Acc \uparrow
Music Models												
MERT-330M [10]	330M	78.6	65.6	61.2	74.7	87.1	76.9	91.3	40.2	72.6	94.4	74.3
MusicFM [11]	330M	83.8	63.9	60.3	76.3	92.0	78.4	91.3	40.0	76.2	91.1	75.3
MuQ [12]	330M	85.6	65.0	62.8	76.1	96.2	81.6	91.4	40.1	79.7	91.2	77.0
Multi-domain Audio Models												
USAD Small [13]	25M	74.1	15.4	48.1	71.8	83.5	74.8	91.0	39.6	73.5	84.8	65.7
USAD Base [13]	97M	73.4	18.7	44.3	71.9	85.8	72.8	91.2	40.0	75.2	88.0	66.1
USAD Large [13]	336M	77.6	29.4	52.6	75.2	85.0	74.2	91.4	40.6	74.9	87.7	68.9
SPEAR Base [15]	94M	82.1	53.2	55.7	75.6	83.8	77.3	91.6	40.7	74.2	88.2	72.2
SPEAR Large [15]	327M	83.1	55.3	62.5	76.8	88.5	78.3	91.8	41.3	76.0	89.7	74.3
SPEAR XLarge [15]	600M	85.5	56.4	62.9	77.9	90.8	78.2	91.8	41.4	77.1	88.8	75.1
Multi-expert Encoder												
Self-supervised (WavLM + ATST + MuQ)												
	734M	85.9	63.5	62.8	76.6	90.2	79.6	91.5	40.7	79.2	91.2	76.1
Supervised (Whisper + AF3)												
	1274M	91.4	43.7	59.9	77.2	78.7	76.4	92.1	42.6	79.6	82.8	72.4
USAD 2.0												
USAD 2.0 Small	25M	73.4	60.2	59.1	78.0	84.3	76.0	91.5	40.7	75.9	90.2	72.9
USAD 2.0 Base	97M	73.8	62.5	58.8	78.8	88.4	77.5	91.7	41.1	76.5	91.8	74.1
USAD 2.0 Large	336M	83.1	62.8	62.2	78.7	89.8	78.5	91.7	41.3	78.7	91.0	75.8
USAD 2.0 XLarge	695M	83.4	62.2	60.5	80.6	89.8	78.4	91.6	41.1	77.7	91.5	75.7
USAD 2.0+												
USAD 2.0 Large+	336M	86.2	56.8	60.9	78.3	90.4	78.3	91.9	41.6	76.8	89.5	75.1
USAD 2.0 XLarge+	695M	86.6	55.1	62.0	78.2	90.3	78.3	91.9	41.7	76.4	89.9	75.0
USAD 2.0 XXLarge+	1036M	87.2	57.9	59.7	80.4	90.1	79.6	91.8	41.4	77.7	90.4	75.6

Table 10: Results on SUPERB [54].

Model	Params	Frame-level			Instance-level			
		PR PER↓	ASR WER↓	SD DER↓	KS Acc↑	IC Acc↑	SID Acc↑	ER Acc↑
Speech Models								
WavLM Base+ [4]	95M	3.9	5.6	3.5	97.4	99.0	89.4	68.7
WavLM Large [4]	317M	3.1	3.4	3.2	97.9	99.3	95.5	70.6
General Audio SSL								
BEATs iter3 [5]	90M	36.4	25.9	5.2	97.7	53.4	57.1	64.5
EAT [6]	88M	55.0	25.9	4.7	92.8	53.7	45.0	62.5
SSLAM [9]	88M	56.4	27.8	4.6	98.8	51.6	42.6	62.6
ATST Frame [8]	86M	20.4	18.8	4.7	95.1	85.4	69.8	64.4
Music SSL								
MuQ [12]	330M	39.9	29.7	5.1	91.5	57.4	49.7	62.3
Multi-domain Audio Models								
USAD Small [13]	25M	7.8	9.5	4.9	96.8	95.5	73.5	66.3
USAD Base [13]	97M	5.1	7.7	4.2	97.1	98.3	88.6	68.0
USAD Large [13]	336M	4.0	6.5	3.9	97.1	98.4	91.2	68.4
SPEAR Base [15]	94M	3.9	3.8	4.1	97.6	98.1	90.0	69.4
SPEAR Large [15]	327M	3.1	3.4	3.8	97.9	99.4	95.0	71.6
SPEAR XLarge [15]	600M	2.9	3.2	3.2	98.1	99.6	96.3	73.3
USAD 2.0 (Self-supervised Teachers)								
USAD 2.0 Small	25M	7.5	9.7	5.4	96.9	95.6	72.5	67.1
USAD 2.0 Base	97M	5.3	8.3	4.8	96.7	96.7	83.6	68.2
USAD 2.0 Large	336M	3.9	5.5	4.8	96.9	95.1	84.7	67.8
USAD 2.0 XLarge	695M	3.6	5.6	4.3	96.2	93.6	90.3	69.1
USAD 2.0+ (Supervised Teachers)								
USAD 2.0 Large+	336M	4.3	5.1	3.6	97.7	99.1	90.1	71.7
USAD 2.0 XLarge+	695M	4.1	5.7	3.6	97.6	99.0	91.6	72.5

	asvspoof 2015	cremad	esc-50	fluent speech	free music archive	fsd50k	fsd kaggle 2018	gtzan	libricount	msynth	speech commands	urban sound 8k	vocal sound	voxceleb	voxlangua 33	aishell-1	clotho	librispeech	meccat	song describer	Track A Avg	Track B Avg
ATST-Frame	0.959	0.395	0.618	0.849	0.660	0.080	0.418	0.828	0.146	0.660	0.922	0.738	0.904	0.976	0.814	0.000	0.412	0.000	0.658	0.502	0.664	0.314
MuQ	0.930	0.383	0.278	0.553	0.573	0.051	0.368	0.778	0.173	0.662	0.908	0.487	0.866	0.309	0.557	0.000	0.278	0.000	0.612	0.488	0.525	0.276
WavLM Base+	0.960	0.337	0.140	0.968	0.553	0.030	0.114	0.626	0.107	0.499	0.959	0.447	0.855	0.518	0.871	0.000	0.243	0.654	0.630	0.353	0.532	0.376
WavLM Large	0.986	0.370	0.293	0.993	0.573	0.037	0.225	0.737	0.189	0.562	0.970	0.486	0.883	0.877	0.901	0.000	0.256	0.038	0.654	0.461	0.605	0.282
Dasheng Base	0.937	0.621	0.755	0.984	0.429	0.063	0.415	0.323	0.386	0.675	0.655	0.829	0.855	0.974	0.311	0.018	0.207	0.103	0.600	0.410	0.614	0.268
Dasheng 0.6B	0.983	0.535	0.593	0.961	0.328	0.041	0.154	0.343	0.087	0.676	0.971	0.701	0.903	0.950	0.561	0.017	0.243	0.000	0.640	0.459	0.586	0.272
Dasheng 1.2B	0.990	0.528	0.627	0.973	0.343	0.043	0.206	0.364	0.172	0.694	0.972	0.760	0.910	0.967	0.585	0.000	0.257	0.000	0.649	0.442	0.609	0.270
USAD Small	0.948	0.320	0.510	0.861	0.630	0.067	0.320	0.798	0.301	0.606	0.957	0.704	0.884	0.297	0.856	0.000	0.387	0.342	0.646	0.411	0.604	0.357
USAD Base	0.927	0.344	0.618	0.954	0.645	0.079	0.337	0.828	0.227	0.621	0.964	0.706	0.891	0.662	0.873	0.000	0.394	0.663	0.650	0.502	0.645	0.442
USAD Large	0.958	0.374	0.513	0.983	0.601	0.064	0.290	0.798	0.219	0.614	0.966	0.667	0.900	0.879	0.891	0.000	0.386	0.741	0.659	0.437	0.648	0.444
SPEAR Base	0.974	0.460	0.443	0.992	0.616	0.067	0.429	0.818	0.116	0.587	0.968	0.729	0.904	0.943	0.863	0.000	0.345	0.504	0.658	0.455	0.660	0.393
SPEAR Large	0.991	0.497	0.523	0.993	0.610	0.075	0.492	0.818	0.271	0.633	0.926	0.759	0.912	0.957	0.907	0.000	0.371	0.668	0.585	0.418	0.691	0.409
SPEAR XLarge	0.990	0.524	0.580	0.994	0.626	0.082	0.563	0.798	0.140	0.624	0.972	0.723	0.910	0.980	0.911	0.000	0.372	0.726	0.676	0.422	0.694	0.439
Whisper Large-v3	0.968	0.561	0.620	0.885	0.619	0.082	0.517	0.828	0.184	0.610	0.947	0.688	0.884	0.974	0.975	0.360	0.374	0.374	0.685	0.491	0.690	0.457
Audio Flamingo 3	0.978	0.665	0.888	0.905	0.878	0.189	0.782	0.919	0.202	0.713	0.943	0.849	0.927	0.970	0.915	0.240	0.448	0.304	0.688	0.452	0.782	0.427
Music Flamingo	0.981	0.625	0.858	0.680	0.828	0.182	0.764	0.889	0.143	0.708	0.893	0.841	0.926	0.977	0.927	0.000	0.452	0.000	0.673	0.502	0.748	0.326
Qwen3Omni	0.972	0.377	0.343	0.955	0.545	0.055	0.459	0.778	0.233	0.577	0.969	0.675	0.883	0.276	0.788	0.000	0.274	0.220	0.652	0.446	0.592	0.319
USAD 2.0 Small	0.948	0.320	0.510	0.861	0.630	0.067	0.320	0.798	0.301	0.606	0.957	0.704	0.884	0.297	0.856	0.000	0.387	0.342	0.646	0.411	0.604	0.357
USAD 2.0 Base	0.927	0.344	0.618	0.954	0.645	0.079	0.337	0.828	0.227	0.621	0.964	0.706	0.891	0.662	0.873	0.000	0.394	0.663	0.650	0.502	0.645	0.442
USAD 2.0 Large	0.968	0.345	0.600	0.920	0.634	0.085	0.410	0.869	0.169	0.637	0.966	0.692	0.895	0.952	0.869	0.000	0.408	0.793	0.664	0.502	0.667	0.473
USAD 2.0 XLarge	0.979	0.361	0.630	0.973	0.649	0.103	0.599	0.869	0.253	0.676	0.954	0.795	0.910	0.977	0.898	0.000	0.419	0.852	0.660	0.496	0.708	0.485
USAD 2.0 Large+	0.977	0.562	0.892	0.992	0.670	0.180	0.758	0.899	0.247	0.693	0.973	0.855	0.922	0.964	0.948	0.408	0.453	0.871	0.694	0.472	0.769	0.580
USAD 2.0 XLarge+	0.970	0.570	0.903	0.991	0.663	0.198	0.790	0.919	0.193	0.697	0.972	0.864	0.932	0.968	0.956	0.515	0.460	0.886	0.696	0.500	0.772	0.611
USAD 2.0 XXLarge+	0.965	0.605	0.918	0.992	0.655	0.191	0.784	0.909	0.346	0.701	0.971	0.861	0.923	0.968	0.954	0.565	0.460	0.881	0.697	0.518	0.783	0.624
WavLM + ATST + MuQ	0.993	0.440	0.422	0.994	0.620	0.058	0.375	0.798	0.135	0.672	0.971	0.529	0.900	0.867	0.897	0.000	0.319	0.827	0.675	0.488	0.645	0.462
Whisper + AF3	0.978	0.732	0.923	0.994	0.900	0.224	0.796	0.869	0.223	0.711	0.976	0.871	0.938	0.985	0.978	0.796	0.473	0.904	0.708	0.543	0.806	0.685

Figure 4: XARES-LLM results on the best-performing audio encoders. Columns 1–15 and 16–20 belong to Tracks A and B, respectively. The last two columns indicate the average scores of Track A and B, respectively. The colors are normalized along each column.

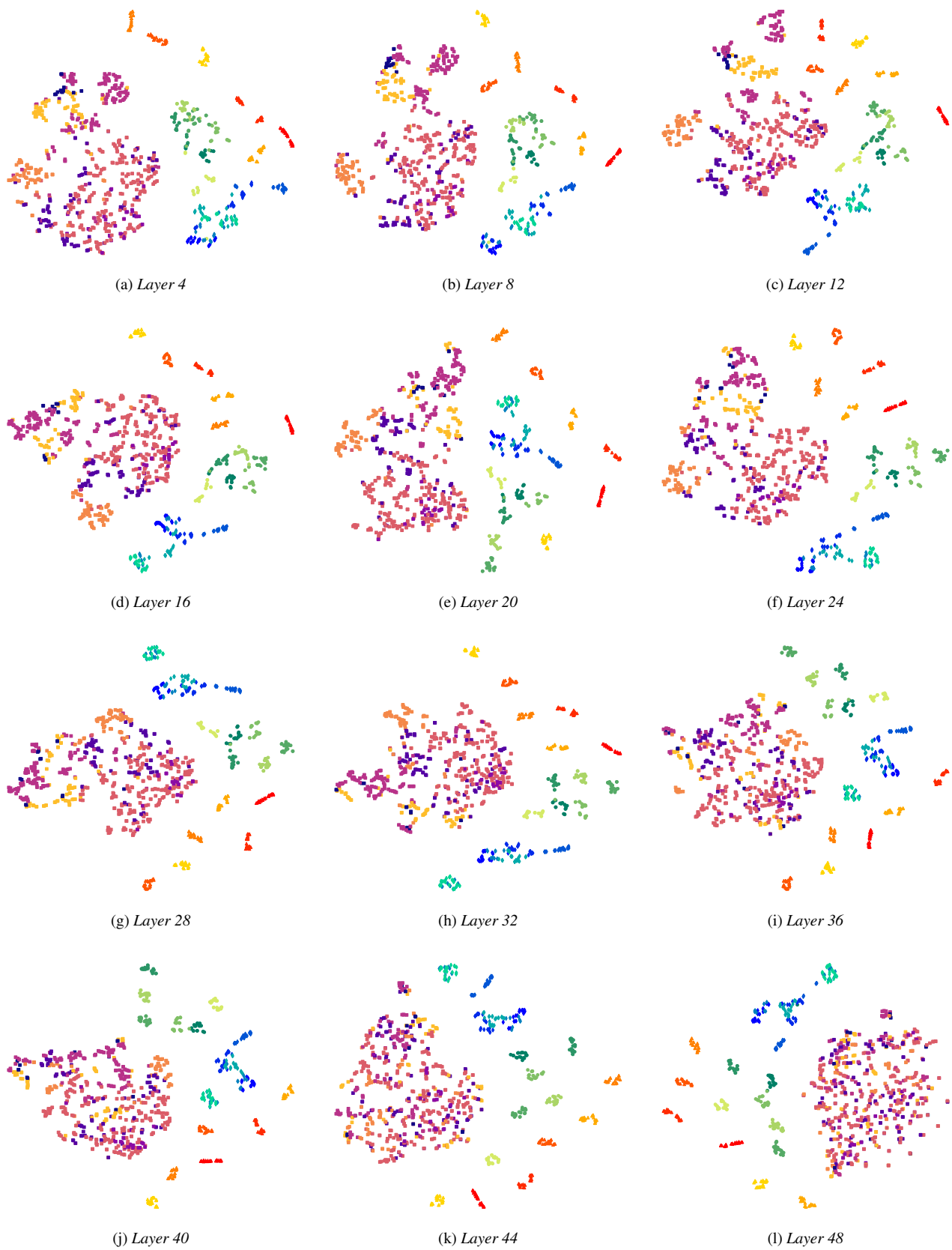


Figure 5: *t*-SNE visualization of USAD 2.0 XXL Large+ hidden representations across the entire model's layers. The legend is shown in Fig. 3.