Stacking Complementary CLAP Embeddings for Improving Text-Audio Alignment Correspondence Scoring

Sheng Li [1] **Jiyi Li** [2] **Takahiro Shinozaki** [1]

Abstract

Text-audio alignment correspondence scoring assesses whether an audio clip satisfies a natural-language prompt by combining semantic grounding, acoustic event recognition, and perceptual quality. We study this problem and compare a wide range of frozen representations, including BEATs, three LAION CLAP variants, MS-CLAP, WavLM, AudioMAE, EAT, Whisper encoder features, and wav2vec2.0 CTC features. A clear pattern emerges: audio-only self-supervised learning (SSL) and speech recognition (ASR) features can improve calibration metrics but do not consistently improve the challenge-ranking metric. By contrast, contrastive audio-language models make complementary errors. We therefore propose CLAP stacking, a lightweight method that combines multiple CLAP embedding spaces. Our final frozen-feature system, evaluated on the scored XACLE test set, improves SRCC from 0.5521 for the best single MS-CLAP pipeline to 0.5934 and reduces MSE from 3.2500 to 2.9418. This performance would fall between official ranks 6 and 7, far above the official baseline SRCC of 0.3345.

1. Introduction

Audio generation and retrieval systems are increasingly evaluated by whether generated or retrieved sound matches a natural-language description. This is more subtle than audio tagging: a clip may contain plausible acoustic events while missing the requested relation, timing, source, or semantic detail. The x-to-audio alignment (XACLE) challenge [1] (XACLE Challenge Organizers, 2026) formalizes this as a regression task, predicting human text-audio alignment correspondence scores for audio-prompt pairs. The target is therefore neither pure acoustic quality nor pure text similarity: it depends on grounded semantic agreement.

We investigate a practical question for such evaluations: given several pretrained audio and audio-language encoders, how should they be combined? A straightforward approach is to choose the strongest backbone and train a regressor. In our experiments, this is not enough. BEATs (Chen et al., 2023) alone captures useful sound-event information but lacks text grounding. CLAP-style models adapt the contrastive language-image pretraining paradigm (Radford et al., 2021) to audio (Elizalde et al., 2023; Wu et al., 2023; Elizalde et al., 2024) and capture audio-text similarity, yet different CLAP models disagree in structured ways because they differ in data, audio frontends, text encoders, and pooling. We exploit this disagreement instead of treating it as noise.

The contributions of this paper are listed as followings. First, we compare contrastive audio-language models, audio-only SSL models, and speech recognition features under the same XACLE pipeline. Second, we show that the strongest improvements come not from adding larger acoustic representations, but from combining complementary CLAP spaces. Third, we introduce CLAP stacking, an effective ensemble inspired by stacked generalization (Wolpert, 1992) that learns when to trust each contrastive space and when to trust cross-model agreement. Another advantage of this approach is that it is relatively lightweight: all encoders are frozen, features are cached, and training uses standard ridge and tree-based regressors.

2. Related Work

Contrastive language-audio pretraining extends the success of CLIP-style image-text learning (Radford et al., 2021) to audio-text pairs. Early CLAP work showed that natural-language supervision can support zero-shot audio understanding (Elizalde et al., 2023). LAION-CLAP scaled this idea with large audio-text data, feature fusion for variable-length audio, and keyword-to-caption augmentation (Wu et al., 2023). MS-CLAP further explored general-purpose audio representations using diverse audio-text supervision

[1]Institute of Science Tokyo, Yokohama, Kanagawa 226-8501, Japan [2]Hokkaido University, Sapporo, Hokkaido 060-0808, Japan. Correspondence to: Sheng Li <sheng.li@ieee.org>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]https://xacle.org

and alternative encoder choices (Elizalde et al., 2024). Our work differs from these pretraining efforts: rather than introducing a new encoder, we study how multiple frozen CLAP spaces can be combined for correspondence scoring.

Audio-only self-supervised learning has also produced strong general-purpose representations, including BEATs acoustic tokenizers (Chen et al., 2023), WavLM for speech processing (Chen et al., 2022), AudioMAE for masked spectrogram modeling (Huang et al., 2022), and EAT for efficient audio transformer pretraining (Chen et al., 2024). Speech recognition models such as Whisper (Radford et al., 2023) and wav2vec2.0 CTC (Baevski et al., 2020) provide another route to semantic evidence through transcription. We evaluate these families as frozen feature sources and find that, for XACLE, they are less effective than stacking audio-language contrastive spaces.

3. Method

3.1. Frozen feature families

For each text-audio pair (x, t), we extract frozen representations from several families. **BEATs** provides 527-dimensional pretrained audio-event logits (Chen et al., 2023). **LAION CLAP HTSAT unfused**, **LAION CLAP HTSAT fused**, and **LAION larger CLAP general** provide normalized contrastive audio and text embeddings (Wu et al., 2023). **MS-CLAP 2024** provides a second 1024-dimensional audio-text embedding space trained with a different recipe and text encoder (Elizalde et al., 2024). We also test audio-only SSL models, including **WavLM Large** (Chen et al., 2022), **AudioMAE** (Huang et al., 2022), and **EAT** (Chen et al., 2024). Finally, we test speech-recognition features from **Whisper-base** (Radford et al., 2023) and **wav2vec2-base-960h CTC** (Baevski et al., 2020).

For a CLAP model with normalized audio embedding a and text embedding b, we build a pair feature

$$\phi(a, b) = [a, b, a \odot b, |a - b|, (a - b)^2, a^\top b]. \quad (1)$$

We concatenate these with BEATs logits. For cross-model features, we include each model's cosine score, absolute differences between cosine scores, low-dimensional SVD projections of each CLAP pair representation, products between these projections, and their absolute differences. These features make the stack sensitive not only to whether a model predicts a good match, but also to whether independent audio-language spaces agree.

3.2. SSL, Whisper, and CTC probes

For WavLM, we pool the final hidden sequence using mean, standard deviation, and max pooling. AudioMAE returns a time-frequency token map; we pool each channel over the token map. EAT is run with its official Kaldi-fbank preprocessing and its CLS/mean/std/max token statistics. Whisper features are mean/std/max-pooled encoder hidden states. The CTC branch uses wav2vec2.0 greedy transcripts, hidden-state pooling, mean and variance of frame confidence, blank ratio, transcript length, and lexical overlap between transcript and prompt. These features are concatenated with the MS-CLAP pair features in probe experiments.

3.3. Stacked regression

Let Φ_m denote a feature family, such as LAION unfused pair features, LAION fused pair features, MS-CLAP pair features, or cross-model agreement features. For each family we train two ridge regressors: one on raw human scores and one on rank-transformed scores. We also train lightweight histogram-gradient regressors on an SVD-compressed concatenation of all features. To avoid leakage, every first-level prediction on the training set is produced out-of-fold using 3 folds. A second-level ridge regressor is then fit on these out-of-fold predictions and applied to averaged first-level predictions on validation or test.

The rank-transformed target is useful because the official ranking metric is Spearman correlation. However, pure rank optimization may worsen MSE. In the final test system, the meta-regressor selected the score target, yielding both strong SRCC and low MSE.

The novelty is not the use of a generic ensemble, but the structure of the ensemble: first-level models are organized around distinct audio-language embedding spaces, and the cross branch explicitly models agreement and disagreement among CLAP variants. This differs from averaging prediction files or concatenating all embeddings into one high-dimensional regressor. Out-of-fold stacking lets the meta-regressor learn when to trust each contrastive space without seeing the held-out labels used to evaluate first-level predictions.

4. Experimental Setup

We use the XACLE train split of 7,500 pairs for model fitting and report on the provided validation and scored test splits of 3,000 pairs each. Metrics are Spearman rank correlation (SRCC), Pearson correlation (LCC), Kendall's tau (KTAU), and MSE; public ranking is by test SRCC. All pretrained models are frozen. Regressors are trained only on XACLE training labels; test labels are used for comparison with the public leaderboard top-ranked systems.

We compare against the official baseline, a BEATs-only baseline, single CLAP pipelines, SSL/audio-only additions, Whisper/CTC additions, and stacked variants. For official leaderboard systems, we cite the public system description provided by the XACLE challenge results page. For MS-CLAP, we use a soundfile-based audio loader to avoid

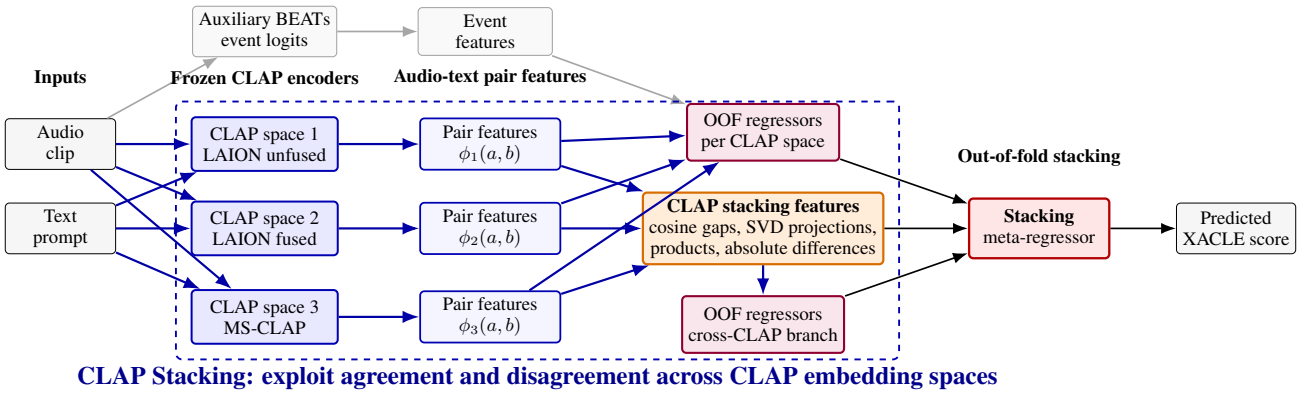

Figure 1. The proposed CLAP stacking pipeline. The central blue path combines three frozen CLAP embedding spaces and explicitly models their agreement and disagreement; BEATs is used only as an auxiliary audio-event feature. First-level regressors are trained out-of-fold (OOF), and a meta-regressor produces the final correspondence score.

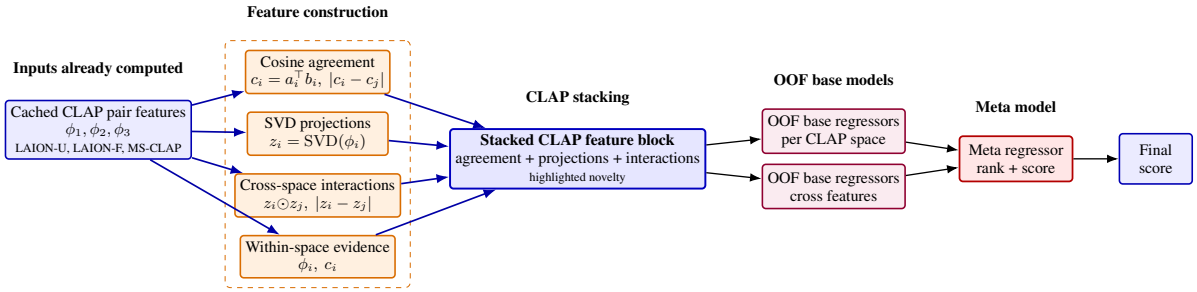

Figure 2. Zoomed-in view of the stacked CLAP stage. The figure starts after frozen CLAP pair features have been cached and focuses on the later part of the method: agreement features, compact SVD projections, cross-space interactions, out-of-fold base regressors, and the final meta-regressor.

Windows torchaudio/FFmpeg issues while preserving 16 kHz resampling. For LAION CLAP we test unfused, fused, and larger-general variants. For Whisper and wav2vec2.0 CTC probes, we use Whisper-base encoder pooling and wav2vec2-base-960h greedy CTC transcripts plus confidence features.

5. Results

Table 1 shows the official top-10 teams, the official baseline, and our ablations in a common test-set view. The single best backbone is MS-CLAP, reaching 0.5521 SRCC. Stacking LAION unfused CLAP with MS-CLAP raises SRCC to 0.5852. Adding the third CLAP, LAION fused, further raises SRCC to 0.5934 and improves MSE to 2.9418. On the public ranking scale, this system would be between the official sixth and seventh teams.

Table 2 gives a broader comparison. Among single CLAP variants, LAION unfused and MS-CLAP are close on validation, while LAION fused and larger-general are weaker. Nevertheless, fused CLAP improves the stack, suggesting that single-model strength is not the only criterion for ensemble usefulness. Cross-model agreement regressors obtain stronger out-of-fold SRCC than individual CLAP regressors, supporting the central hypothesis that disagreement between CLAP spaces is informative.

The progression from single models to stacks is consistent across validation and test. The 2-CLAP stack improves test SRCC by 0.0331 over MS-CLAP; the 3-CLAP stack adds another 0.0082. While this second increment is smaller, it also improves LCC and MSE, indicating that fused CLAP contributes useful calibration rather than only perturbing ranks.

The self-supervised learning (SSL) and speech recognition (ASR) probes show a different pattern. WavLM Large, AudioMAE, and EAT do not outperform MS-CLAP-only on the ranking metric when directly concatenated. Whisper+CTC improves LCC, KTAU, and MSE in the 512/128 probe but slightly lowers SRCC relative to MS-CLAP-only. Inspection of CTC transcripts shows many empty or fragmentary outputs, consistent with the hypothesis that XACLE contains many non-speech or weakly speech-like audio clips. Thus, ASR evidence is useful for calibration when speech is present, but not a robust primary alignment signal for the whole dataset.

Table 1. XACLE test-set comparison with the official team ranking, official baseline, and our ablations. The official leaderboard is ranked by SRCC.

Rank / Type	System	SRCC	LCC	KTAU	MSE	Note
1	Sun_NPU_01 (Sun et al., 2026)	.6402	.6873	.4612	3.0111	official
2	Chunarkar_NTHU_01 (Chunarkar et al., 2026)	.6382	.6851	.4596	2.8256	official
3	Shiota_TMU_4 (Tsutsumi et al., 2026)	.6327	.6426	.4564	9.5999	official
4	Huang_WHU_01 (Liu et al., 2026)	.6264	.6695	.4497	2.8369	official
5	Guan_HEU_02	.6143	.6770	.4403	2.8044	official
6	Takano_UTokyo_03 (Takano & Yoshida, 2026)	.6142	.6542	.4407	2.9850	official
Ours	**3-CLAP stack**	**.5934**	**.6507**	**.4233**	**2.9418**	**would rank 7**
7	Huy_HCMUS_01	.5702	.6232	.4181	3.9974	official
8	Nakagome_LINEWORKS_01	.5679	.6262	.4038	3.2625	official
9	Niizumi_TMUNTT_3 (Niizumi et al., 2026)	.5672	.6346	.4021	3.9274	official
10	Kumar_TII_1 (Kumar et al., 2026)	.5624	.6205	.3979	3.2195	official
Ours	2-CLAP stack	.5852	.6396	.4165	3.0153	unfused LAION + MS-CLAP
Ours	MS-CLAP single	.5521	.6035	.3901	3.2500	best single model
Ours	LAION CLAP single	.5361	.5842	.3784	3.3782	unfused HTSAT
Ours	LAION fused CLAP single	.5083	.5530	.3570	3.5872	fused HTSAT
Official	Baseline (XACLE Challenge Organizers, 2026)	.3345	.3420	.2290	4.8110	organizer baseline
Ours	BEATs only	.2584	.2686	.1747	5.7734	audio-event logits

Official systems are cited only when the XACLE results page provides a public system-description PDF; Guan_HEU_02, Huy_HCMUS_01, and Nakagome_LINEWORKS_01 list no paper link in the official JSON.

Table 2. Validation and probe comparisons. Full validation uses all 7,500 training and 3,000 validation pairs. Probe results use 512 training and 128 validation pairs, except the EAT smoke test which uses 256/64.

System	Split	SRCC	LCC	KTAU	MSE
3-CLAP stack	full	**.6476**	**.6718**	**.4698**	7.1189
2-CLAP stack	full	.6439	.6671	.4664	7.0548
CLAP rank ensemble	full	.6061	.6357	.4358	7.0520
LAION unfused CLAP	full	.5988	.6286	.4302	**3.4297**
MS-CLAP	full	.5979	.6135	.4288	3.5347
LAION fused CLAP	full	.5631	.5926	.4016	3.6744
LAION larger-general CLAP	probe	.4858	.5180	.3468	4.3842
MS-CLAP + WavLM Large	probe	.5619	.6140	.4062	3.7158
MS-CLAP + AudioMAE	probe	.5452	.5716	.3879	4.0314
MS-CLAP + EAT	smoke	.4956	.5978	.3679	4.9569
MS-CLAP + Whisper + CTC	probe	.5676	.6393	.4165	3.5478
MS-CLAP only	probe	.5712	.6002	.4107	3.8371
Official baseline	full	.3844	.3961	.2646	4.8361

6. Discussion

CLAP stacking is attractive for workshop-scale systems because it is simple, reproducible, and computationally modest after feature extraction. It also gives interpretable diagnostics: single-model CLAP scores capture direct alignment, while cross-model agreement captures reliability. The main limitation is that all encoders remain frozen and the stack is trained only as a regressor. Future work should study rank-aware losses, calibration that preserves SRCC while improving MSE, and lightweight fine-tuning of CLAP projection heads on human correspondence judgments.

7. Conclusion

We presented a frozen-feature CLAP stacking method for text-audio alignment correspondence scoring. Across extensive baselines on XACLE, adding more audio-only SSL or ASR features was less effective than stacking complementary CLAP embedding spaces. A three-CLAP stack combining BEATs, LAION CLAP unfused, LAION CLAP fused, and MS-CLAP achieves 0.5934 test SRCC, substantially improving over the official baseline and the best single CLAP pipeline.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech

representations. *Advances in Neural Information Processing Systems*, 2020.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Zeng, M., and Wei, F. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., and Wei, F. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning*, 2023.

Chen, W., Liang, Y., Ma, Z., Zheng, Z., and Chen, X. EAT: Self-supervised pre-training with efficient audio transformer. *arXiv preprint arXiv:2401.03497*, 2024.

Chunarkar, S. B., Hamza, K., and Lee, C.-C. Cross-modal semantic alignment via ensemble audio-text features for XACLE challenge. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Chunarkar_NTHU_Chunarkar_NTHU.pdf, 2026.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. CLAP: Learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

Elizalde, B., Deshmukh, S., and Wang, H. Natural language supervision for general-purpose audio representations. *arXiv preprint arXiv:2309.05767*, 2024.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. AudioMAE: Masked autoencoders are efficient learners for self-supervised audio representation learning. *Advances in Neural Information Processing Systems*, 2022.

Kumar, G. K., Lepauloux, L., and Hacid, H. WavLink Score: Audio-text semantic alignment scoring system using dual-encoder features and CatBoost regression. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Kumar_TII_Kumar_TII.pdf, 2026.

Liu, Z., Liu, S., Yang, X., Lyu, S., and Huang, G. EnTA-Align: Ensemble-driven text-audio alignment. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Huang_WHU_Huang_WHU.pdf, 2026.

Niizumi, D., Takeuchi, D., Yasuda, M., Nguyen, B. T., Harada, N., and Ono, N. Trust your CLAP: TMU-NTT XACLE challenge 2026 technical report. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Niizumi_TMUNTT_Niizumi_TMUNTT.pdf, 2026.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *Proceedings of the International Conference on Machine Learning*, 2023.

Sun, B., Yin, H., and Xiao, Y. TASALIGN: An objective evaluation method for text-audio semantic alignment based on mixture of experts and sequential cross-attention. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Sun_NPU_Sun_NPU.pdf, 2026.

Takano, T. and Yoshida, R. Technical report: The university of tokyo XACLE challenge submission. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Takano_UTokyo_Takano_UTokyo.pdf, 2026.

Tsutsumi, A., Tanaka, K., and Shiota, S. The TMU submission to the XACLE challenge at ICASSP 2026 SP grand challenge. XACLE Challenge 2026 system description, https://xacle.org/docs/paper/Shiota_TMU_Shiota_TMU.pdf, 2026.

Wolpert, D. H. Stacked generalization. *Neural Networks*, 5 (2):241–259, 1992.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Nezhurina, M., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *arXiv preprint arXiv:2211.06687*, 2023.

XACLE Challenge Organizers. The first XACLE challenge. https://xacle.org/results.html, 2026.