TRACE: Training-Free Partial Audio Deepfake Detection via Embedding Trajectory Analysis of Speech Foundation Models

Supplementary Material

This supplementary material provides four additional analyses that support and extend the results in the main paper: (A) the complete feature statistics ranking across all 43 statistics on PartialSpoof; (B) the encoder layer ablation showing why layer 18 of WavLM-Large is optimal; (C) the score orientation stability analysis confirming that minimal calibration data is required; and (D) extended discussion of key findings and future directions.

6. Complete Feature Statistics Ranking

Figure 6 reports the EER of all 43 evaluated statistics on the PartialSpoof evaluation set using WavLM-Large layer 18. Statistics are grouped into five families: base F1 (rms, std, mean-abs, kurtosis), multi-scale derivatives (dt2/3/4), sliding-window variants (max, min, spread), percentile and tail statistics, and directional angle statistics. F2 statistics are omitted from this ranking as all score approximately 50% EER at layer 18, carrying no discriminative information, as confirmed in the main ablation.

The key observations are as follows. First, **F1-rms** (11.07%), **F1-mean-abs** (10.84%), and **F1-dt4-rms** (11.08%) are the three strongest single statistics, with performance clustered within 0.25 percentage points of each other. Second, multi-scale derivative statistics (F1-dt2 through F1-dt4) perform comparably to base statistics, confirming that splice-induced onset patterns are visible at multiple temporal scales. Third, sliding-window variants (F1-maxW, F1-p99) rank lower in-dataset but provide complementary information for cross-domain generalization, as discussed in the main paper. Fourth, directional angle statistics (`angle_mean`, `angle_rms`, `angle_std`) perform poorly standalone (22–50% EER) but improve cross-lingual transfer when combined with magnitude statistics, motivating their inclusion in the $F1_{opt}$ combination for HAD. Fifth, kurtosis-based statistics are unstable due to their sensitivity to outliers, consistent with the encoder ablation in Figure 4 of the main paper.

7. Encoder Layer Ablation

Figure 7 shows the per-layer EER of WavLM-Large on PartialSpoof using the F1-rms statistic, sweeping across all 24 transformer layers. The results reveal a clear and consistent pattern. EER is highest at the final layer (layer 24, 62.0% with F1-rms) and decreases steadily through intermediate layers, reaching a minimum at **layer 18** (11.07% EER). Performance then degrades again at shallower layers (below layer 12), where representations are too low-level to reliably capture phonological transitions.

Table 6. Optimal layer and EER per encoder (PartialSpoof, F1-rms statistic). WavLM models benefit most from intermediate-layer extraction.

Encoder	Optimal layer	EER (%)
WavLM-Large	18	11.07
WavLM-Base	10	12.43
HuBERT-Large	16	14.82
Wav2Vec2-XLSR	20	19.31
Wav2Vec2-Base	9	21.14
Whisper-Base	4	27.83

This pattern has a clear mechanistic interpretation. The final layer of WavLM-Large is explicitly trained to predict discrete speech units, a semantic-level objective that creates smooth, averaged representations where frame-level acoustic discontinuities are suppressed. Layer 18 lies just past the phoneme-encoding peak identified in prior probing studies: it captures fine-grained acoustic transitions while retaining enough structure for the F1 sequence to be meaningful. This suggests that the discriminative window for splice detection corresponds precisely to the phonological representation layer, not the semantic output layer, an insight that may guide future work on foundation model-based acoustic forensics.

Table 6 reports the layer-wise EER for all six encoders at their optimal single-statistic configuration. WavLM-Large consistently benefits most from intermediate-layer extraction, while Whisper-Base shows a different profile reflecting its ASR-oriented training objective.

7.1. First-Order vs Second-Order Dynamics

Table 7 compares first-order (F1) and second-order (F2) dynamics across encoders and statistics on PartialSpoof. F1 consistently and substantially outperforms F2 across every encoder and statistic combination. The best F1 result (WavLM-Large, F1-std: 16.37% EER) outperforms the best F2 result (Whisper-Base, F2-mean-abs: 27.83% EER) by over 11 percentage points. This gap is structural: F1 directly measures the magnitude of the embedding displacement at each frame transition, producing a sharp spike at splice boundaries. F2, by contrast, mea-

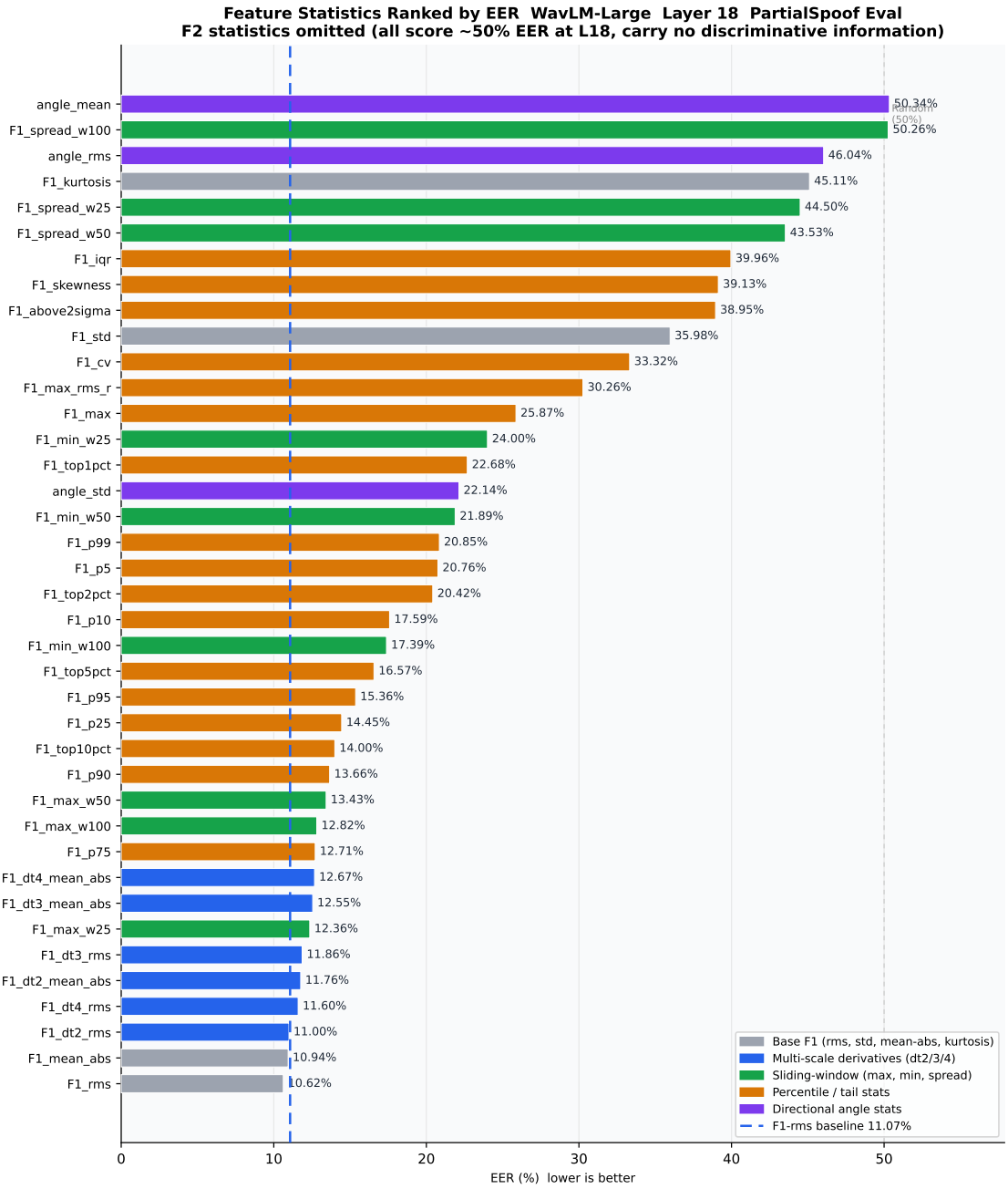

Figure 6. EER of all 43 feature statistics on PartialSpoof (WavLM-Large, layer 18), ranked from best to worst. F1-rms, F1-mean-abs, and F1-dt4-rms perform best (10.84–11.07%). Directional angle features are weak standalone but aid cross-domain generalization when fused with magnitude statistics. All F2 statistics score ≈50% EER and are omitted.

sures the *rate of change* of that displacement, which is informative only if splice boundaries have characteristic entry and exit ramps a pattern not observed in the data. As a result, F2 collapses near chance at the optimal encoder layer across all configurations. The advantage of F1 is largest for WavLM-Large (11.3 pp gap), whose denoising masked prediction objective preserves temporal structure more faithfully than the ASR-oriented Whisper-Base (5.8 pp gap), consistent with the encoder analysis in the main paper.

8. Extended Discussion

8.1. Why the LlamaPS result matters

The LlamaPartialSpoof results carry a particularly important message. This benchmark uses LLM-driven synthesis tools, including ElevenLabs and comparable commercial systems, that produce outputs of unprecedented

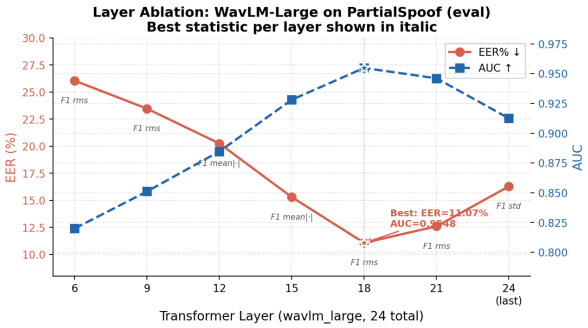

Figure 7. Per-layer EER of WavLM-Large on PartialSpoof (F1-rms statistic). Layer 18 achieves the minimum EER (11.07%). The final layer (layer 24) performs near chance due to semantic-level smoothing of acoustic discontinuities.

Table 7. First-order (F1) vs second-order (F2) dynamics across encoders and statistics on PartialSpoof eval. F1 consistently outperforms F2 across all combinations, confirming that embedding transition rate carries the dominant forensic signal in frozen speech foundation model representations.

Encoder	Feature	EER (%)↓	AUC↑
First-order dynamics (F1)			
WavLM-Large	F1-std	**16.37**	**0.912**
HuBERT-Large	F1-std	22.88	0.852
Whisper-Base	F1-mean-abs	29.49	0.766
Whisper-Base	F1-rms	29.73	0.765
Whisper-Base	F1-std	33.24	0.725
WavLM-Base	F1-std	37.56	0.672
Wav2Vec2-Base	F1-mean-abs	42.04	0.617
Second-order dynamics (F2)			
Whisper-Base	F2-mean-abs	27.83	0.798
WavLM-Large	F2-std	27.66	0.787
WavLM-Large	F2-rms	27.67	0.787
HuBERT-Large	F2-kurtosis	31.91	0.737
WavLM-Large	F2-kurtosis	35.50	0.695
HuBERT-Large	F2-std	36.64	0.676
HuBERT-Large	F2-rms	36.66	0.675
WavLM-Large	F2-mean-abs	38.04	0.663
Whisper-Base	F2-rms	39.03	0.665
Whisper-Base	F2-std	39.04	0.665
WavLM-Base	F2-std	41.56	0.616
WavLM-Base	F2-rms	41.59	0.616
Wav2Vec2-Base	F2-mean-abs	40.04	0.633

naturalness. These are the tools being actively misused for disinformation, voice cloning, and audio fraud. Supervised detectors trained on older-generation PartialSpoof data show near-random performance on LlamaPS (35–47% EER), as confirmed by multiple recent studies. TRACE, trained on nothing, achieves 24.12% Free EER and 19.82% EER on partial-fake subsets, outperforming every published supervised baseline on this benchmark. The implication is that foundation model dynamics generalize to unseen synthesis technology in a way that task-specific fine-tuning does not, and that the forensic signal we exploit is encoding-technology agnostic: present whether the fake segment was produced by a unit-selection system, a flow-based model, or an LLM-driven synthesis engine.

8.2. Why HAD and ADD 2023 are harder

The relatively higher EERs on Mandarin benchmarks (HAD 20.92%, ADD 2023 33.43%) should not be overinterpreted as a language barrier. Our argument for language independence is supported by the fact that a system calibrated entirely on English (PartialSpoof) detects Mandarin fakes well above chance. The primary difficulty is spoof **segment length**: HAD and ADD 2023 contain shorter, more densely packed spoof segments whose F1 spike is diluted by global score aggregation. The sliding-window statistic $F1_{\max W}$ partially recovers this signal (HAD: 30.11% $\rightarrow$ 20.92%). Future work on frame-level anomaly maps should close this gap further.

8.3. Fully-fake utterances: a principled scope constraint

TRACE is designed to detect splice boundaries and is not expected to detect end-to-end TTS utterances, which have no such boundaries. The LlamaPS results confirm this precisely: partial-fake EER (13–16%) is strong while fully-fake EER ($\approx$45%) is near-chance. This is not a design flaw but a scope constraint. Future work could combine TRACE with a complementary fully-fake detector based on spectral artifacts or prosodic consistency for a comprehensive training-free detection pipeline.

8.4. Broader implications for foundation model research

Our work suggests a clear answer to what is next in multimodal foundation models: **behavioral analysis of embedding dynamics**. Rather than fine-tuning massive models on ever-larger labeled datasets, we can interrogate the internal geometry of frozen representations to answer forensic questions cheaply, without gradients, and without labeled fake data. Several directions remain open: frame-level anomaly maps could enable segment-level localization, directly addressing the short-spoof-segment weakness on HAD and ADD 2023; multi-layer fusion across layers 15–21 may improve robustness beyond the single optimal layer; and the same paradigm could extend beyond audio to deepfake face detection via vision transformers, machine-generated text detection via language models, or cross-modal consistency verification in multimodal foundation models.