Supplementary Material

CAST at SignEval 2026: Channel-Aware Spatial Transfer Learning with Pseudo-Image Radar for Sign Language Recognition

This document provides additional quantitative and qualitative analysis supplementing the main paper. All results are computed on the 10% holdout validation split (`StratifiedShuffleSplit`, `random_state=42`, N=1386 samples) used during training, achieving an overall Top-1 accuracy of 84.6% on this split.

S1 Confusion Matrix Analysis

Figure S1 shows a 16×16 sub-matrix extracted from the full 126×126 confusion matrix. The 16 classes are chosen by identifying all off-diagonal entries with the largest raw error counts, then taking the union of the class indices involved until 16 distinct classes are selected. This procedure surfaces the tightest inter-class clusters and avoids masking localized confusion by averaging over the full vocabulary. Two systematic failure clusters are visible.

Short-duration gestures (left block, rows/columns with high off-diagonal counts): Gestures with fewer than approximately 15 slow-time frames (≈1 s at 13 fps) generate cadence spectra dominated by Blackman–Harris window sidelobe artifacts rather than genuine periodic peaks. The Cadence Velocity Diagram (CVD) adds noise rather than a discriminative signal in these cases, and the cross-attention fusion propagates that noise into the fused representation.

Finger-spelled alphabet confusion (right block): Several alphabet letters share nearly identical gross-motion profiles. At typical signer distances of 1–2 m, the 60 GHz wavelength ($\lambda \approx 5$ mm) is comparable to inter-finger gaps, so the Range-Time Map (RTM) cannot resolve fine static finger configurations. Because no differential cadence exists between signs sharing the same hand shape, the CVD is equally uninformative. These pairs account for a disproportionate share of total errors. Together, these two clusters account for approximately 30% of all validation errors. They represent a ceiling imposed by sensor physics and capture rate, not by the classifier design and both are explicitly discussed in Section 4.6 of the main paper.

S2 RTM and CVD Map Comparisons for Confused Pairs

Figure S2 examines in parallel the RTM and CVD maps of misclassified samples (columns 1–2) against correctly-classified representative samples from the predicted class (columns 3–4) for the four most frequently confused class pairs. Several patterns are apparent. First, in the short-duration failure cases, the RTMs of the confused pair are visually similar in temporal extent and gross energy profile, while the CVDs show broad, flat spectra with no clear cadence peak in either. Second, in the finger-spelling cases, the RTMs are virtually indistinguishable at the spatial resolution available in magnitude-only data. The CVDs likewise carry no differential cadence information. These examples support the physical argument in the main paper: the observed errors are structurally unresolvable from magnitude RTM data alone, and adding phase information or higher-range resolution inputs would be required to break the ceiling.

S3 Ablation: CASA Head Count

The main paper (Table 2) reports only the 4-head CASA configuration because it was used in the final submission. Table S1 provides a more complete evaluation using the same 5-fold CV protocol.

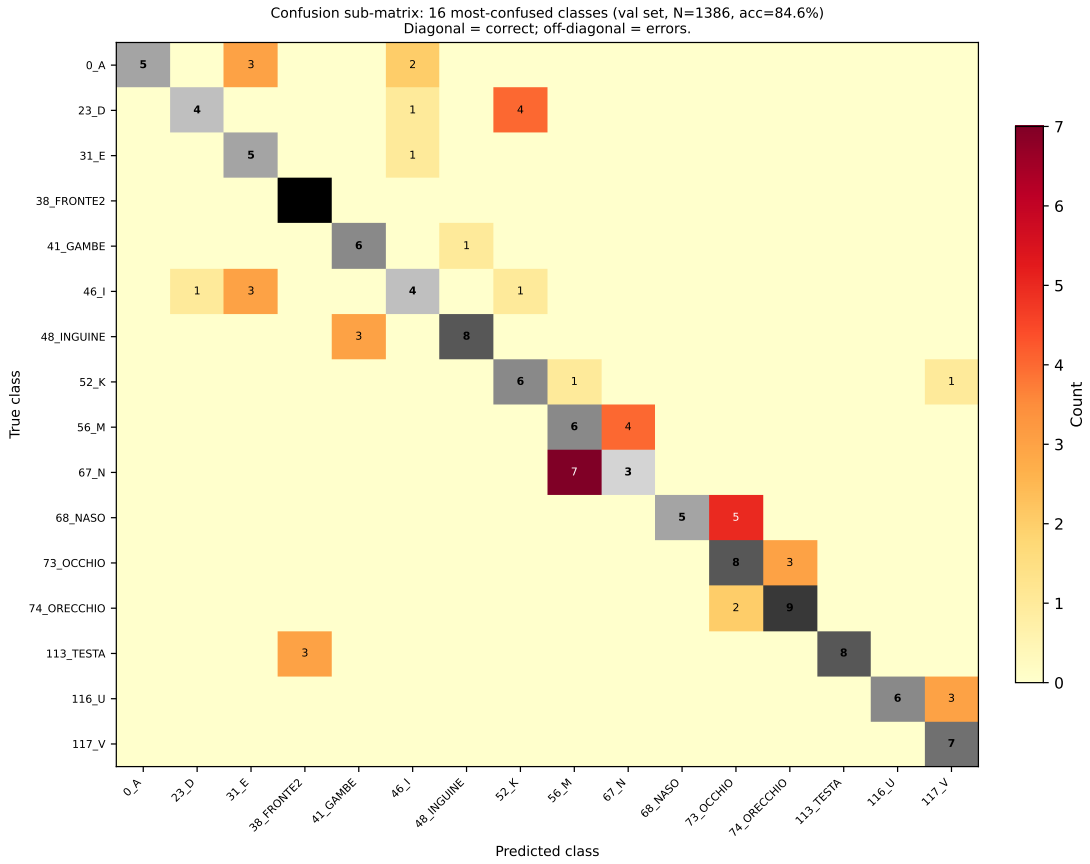

Figure S1: Confusion sub-matrix: 16 most-confused classes on the 10% validation set (N=1386). Diagonal cells (grey) show correct predictions; off-diagonal cells (orange scale) show errors. Row label = true class; column label = predicted class. Class names are truncated to 12 characters; full names follow the MultiMeDaLIS dataset convention (`id_sign_name`).

Table S1: Effect of CASA head count on validation accuracy. All other hyperparameters are identical. The small gap between 1-head and 4-heads configurations (+0.3%) confirms the discussion in Section 3.2 of the main paper with N=3 antenna tokens. The added capacity of multiple heads is modest.

CASA configuration	5-fold Acc (%)	Δ vs. no CASA
Remove CASA (3-channel stacking)	79.8 ± 1.1	—
CASA, 1 head	80.2 ± 1.0	+0.4
CASA, 4 heads (ours)	$\mathbf{80.5 \pm 0.9}$	+0.7

Note: These single-model 5-fold CV numbers differ slightly from the competition Kaggle scores reported in the main paper, which reflect a single 90/10 run with a 7-checkpoint ensemble. See Section 4.1 of the main paper for the full evaluation protocol.

S4 Per-Class Accuracy Breakdown

Figure S3 shows the complete per-class Top-1 accuracy for all 126 classes, sorted in ascending order. Classes are not resolvable from magnitude RTM data due to sensor-physics ceilings (short duration / finger-spelling, denoted (†) in the figure caption) are concentrated in the red (<50%) and orange (50–70%) bars on the left, confirming the structural failure analysis of Section 4.6 of the main paper.

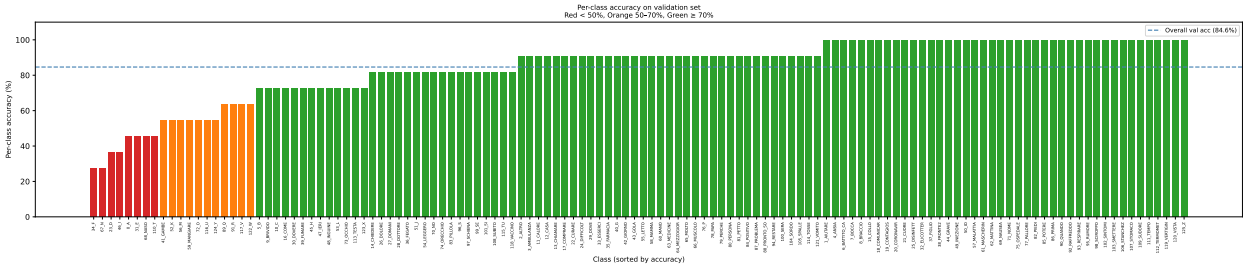

Figure S3: Per-class Top-1 accuracy for all 126 classes on the 10% validation split (N=1386, overall accuracy = 84.6%). Bars are colour-coded: **red** $< 50\%$, **orange** 50–70%, **green** $\geq 70\%$. The dashed blue line marks the overall mean of 84.6%. Classes below 50% are predominantly finger-spelled alphabet letters (A, E, Y, S, ...)$^{(\dagger)}$ whose static hand shape differences are unresolvable at 60 GHz $\lambda \approx 5$ mm without phase data, and short-duration gestures whose CVD is sidelobe-dominated rather than cadence-resolved.

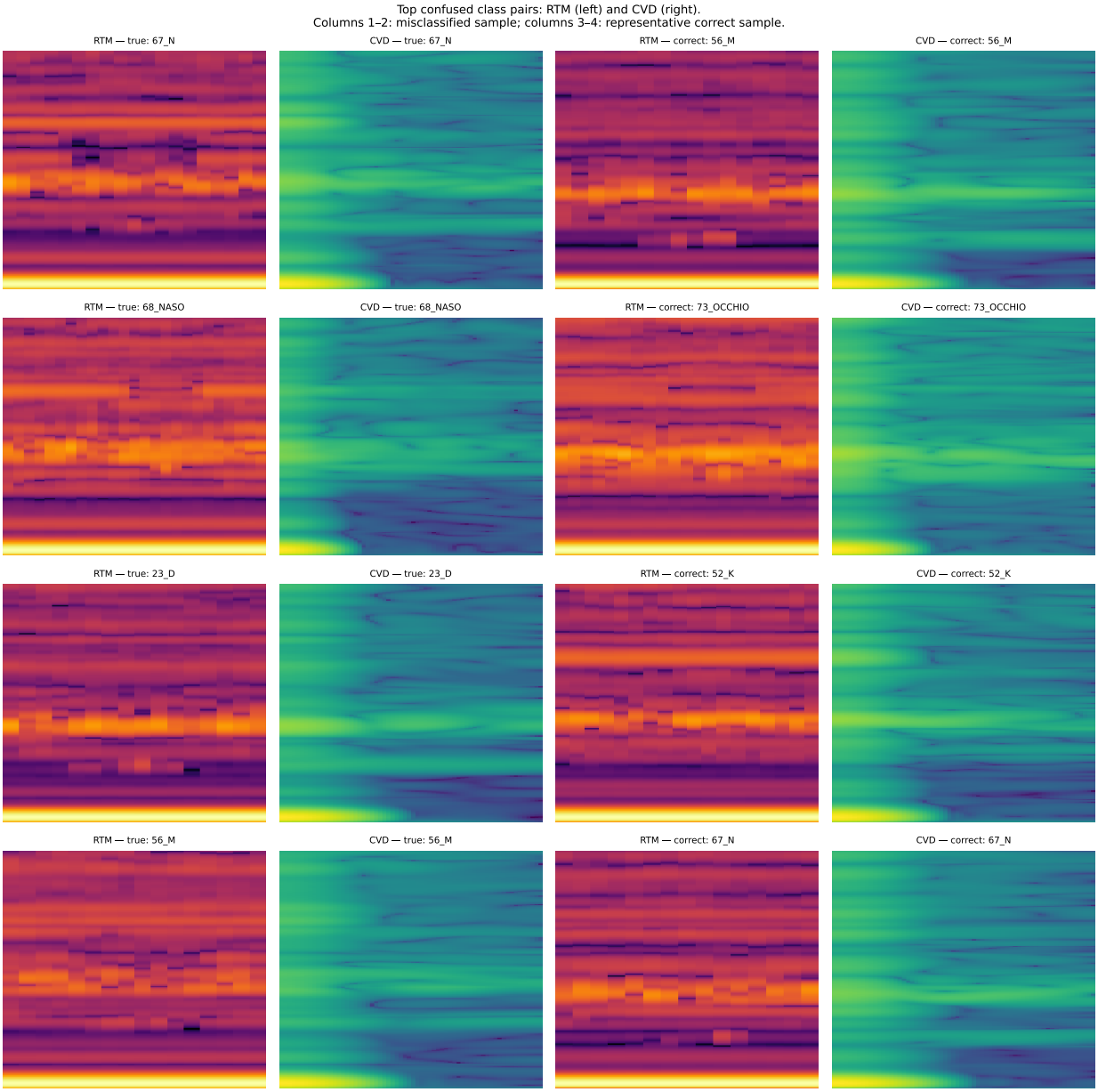

Figure S2: Top-4 confused class pairs: RTM and CVD maps. Each row corresponds to one confused pair (true class $\rightarrow$ predicted class). *Column 1:* RTM of a misclassified sample (true class). *Column 2:* CVD of the same sample. *Column 3:* RTM of a correctly-classified sample from the predicted class. *Column 4:* CVD of the same correct sample. The row label shows the number of such errors in the validation set. Antenna 1 is shown in all panels.

S5 Training Hyperparameters

Table S2 consolidates all training hyperparameters for reproducibility. All values apply to the CAST full model.

Table S2: Complete training hyperparameter listing for CAST.

Hyperparameter	Value
Optimiser	
Optimiser*	AdamW
Learning rate* (EffNetV2-S baseline)	2×10^{-4}
Learning rate* (ConvNeXt-T baseline)	3×10^{-4}
Learning rate (CAST)	3×10^{-4}
Weight decay*	0.05
Gradient clip (max norm)	1.0
Schedule	
Total epochs (CAST)	70
Total epochs (baseline)*	45
Warmup epochs	5
LR schedule	Cosine annealing
Min. LR fraction	0.01
Regularisation and augmentation	
Label smoothing $\epsilon_{\text{ls}}{}^\star$	0.1
Auxiliary loss weight λ_{aux}	0.3
MixUp $\alpha^\star$	0.4
CutMix $\alpha^\star$	1.0
SpecAugment masks (freq/time)	≤ 2 / ≤ 2
Temporal warp σ	0.15
Magnitude warp σ / knots	0.1 / 4
Multipath delay / attenuation	≤ 10 bins / 5–15%
Antenna dropout probability	0.1
Stochastic Weight Averaging / EMA	
SWA start epoch	56
SWA start epoch (baseline)*	36
EMA decay	0.9995
Input and batch	
Input spatial size	224×224
Max temporal length T_{max}	48 frames
Batch size (total)	48
Precision	AMP (fp16)
Inference ensemble	
Checkpoint ensemble size	7 (top-5 + EMA + SWA)
TTA views per sample	5

$\star$ denotes values shared with the single-backbone baseline.