Supplementary materials

1 Segmentation decoder outputs

To complement the quantitative results in the main paper, this section visualises the segmentation decoder predictions obtained from the best checkpoint of each leave-one-dataset-out protocol. For every protocol we randomly sample test frames from the held-out set **O**, **C**, **M** or **I** and additionally from the common out-of-distribution split **S**. Each image triplet shows the RGB input, the predicted binary spoof mask, and the mask overlaid on the input. Dataset abbreviations follow the convention introduced in the paper: **O** – *OULU-NPU*, **C** – *CASIA-FASD*, **M** – *MSU-MFSD*, **I** – *Idiap Replay-Attack*, **S** – *SiW-Mv2* (out-of-distribution).

Figure 1 illustrates how our leading model-chosen from three independent training runs-responds to four CASIA-FASD test frames. The two paper-based attacks (*Cutout paper*, *Printed paper*) trigger a uniform red mask across the entire face, while the *Screen* replay induces two distinct red clusters at the lower facial region. By contrast, the bona fide frame remains completely unmarked, confirming that the network refrains from false positives on genuine inputs.

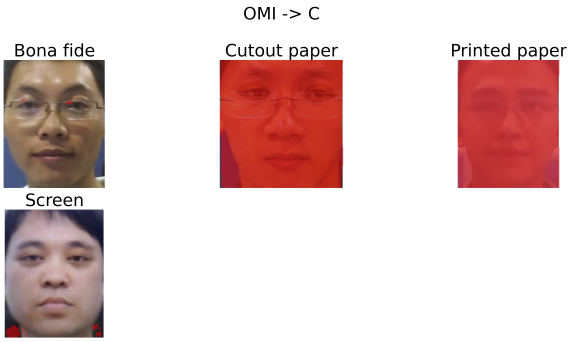

Figure 1: Protocol **OMI→C**: random CASIA-FASD frames with predicted spoof masks.

Turning to Replay-Attack (Figure 2), the model exhibits a different failure pattern: it erroneously marks the bona fide image as spoof, yet completely misses the paper-print attack, leaving its mask empty. Happily, the genuine screen-replay case is correctly identified, with full-face spoof coverage.

In Figure 3, the OULU-NPU examples reveal near-perfect behavior: the genuine face produces a blank mask, and both printed-paper and screen-replay attacks are uniformly highlighted in red. This consistent separation underscores the model's effectiveness when trained on combined Contrastive + MSE losses.

Figure 2: Protocol **OCM→I**: qualitative results on Idiap Replay-Attack.

Figure 3: Protocol **CMI→O**: held-out OULU-NPU frames with spoof overlays.

The MSU-MFSD test set (Figure 4) again demonstrates the model's reliability: bona fide inputs yield empty masks, while both printed-paper and screen-replay attacks trigger full-face spoof masks. Such uniform detection showcases the approach's stability across datasets.

Figure 4: Protocol **OCI→M**: spoof masks on MSU-MFSD frames.

Finally, when challenged with out-of-distribution SiW-Mv2 samples (split **S**), the model maintains its discriminative power. As shown in Figs. 5–8, the bona fide example remains unmasked, while all 15 novel attack types-from cosmetics and obfuscation to various masks and replay attacks-are consistently and comprehensively highlighted. These results confirm robust generalization to unseen presentation manipulations.

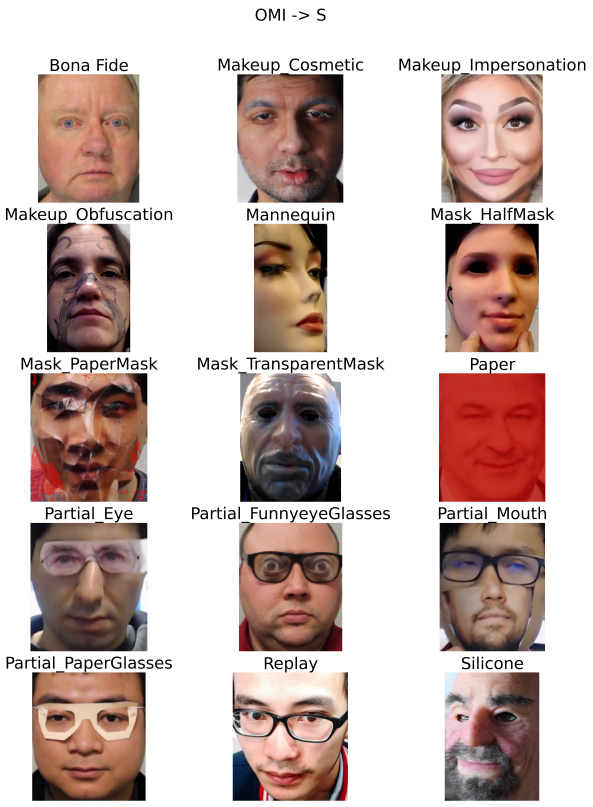

Figure 5: Out-of-distribution evaluation (**S** = SiW-Mv2) for model trained under OMI→C.

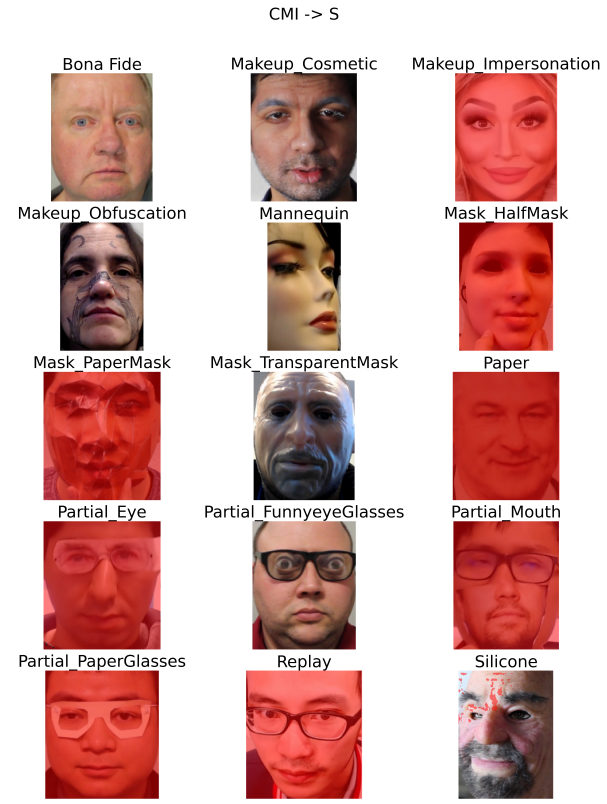

Figure 7: SiW-Mv2 (**S**) results for model trained under CMI→O.

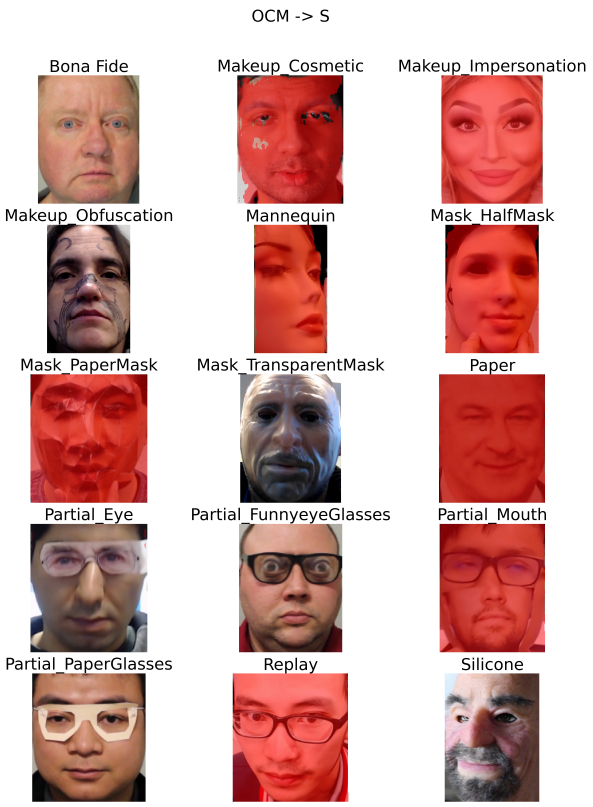

Figure 6: Same model as Figure 2, evaluated on SiW-Mv2 (**S**).

Figure 8: SiW-Mv2 (**S**) results for model trained under OCI→M.

2 Parameters search

We performed a grid search over contrastive- and segmentation-loss weights while holding the classification-loss coefficient fixed at 1.0 (Table 1). In addition, we varied the decoder's output dimensionality by using either two channels (binary live/spoof mask) or three channels (separate spoof-type classes), to assess both binary and multi-class segmentation setups. In Table 1, the best mean HTER scores are highlighted by color (red = CASIA-FASD, blue = Idiap Replay-Attack, green = MSU-MFSD, yellow = OULU-NPU). Across nearly all datasets, higher segmentation-loss weights (>0.2) yielded superior performance, confirming that the segmentation component is crucial to training. By contrast, the optimal contrastive-loss weight varied by dataset, with its best values distributed throughout the search range. Due to the combinatorial explosion of possible weight combinations and our available computational budget, we restricted both contrastive and segmentation weights to the three values 0.2, 0.5, 1.0.

Contrastive Loss Weight	Segmentation Loss Weight	Seg. Output Channels	Dataset	HTER(%)↓	AUC(%)↑	Seg. HTER(%)↓
0.2	0.2	2	casia_fasd	12.11 ± 0.29	94.86 ± 0.63	16.41 ± 2.51
			msu_mfsd	4.76 ± 1.45	98.06 ± 0.42	3.47 ± 2.02
			oulu_npu	8.70 ± 0.23	97.24 ± 0.20	13.28 ± 5.16
			replay_attack	14.67 ± 0.93	87.59 ± 3.00	5.33 ± 1.04
		3	casia_fasd	9.41 ± 1.12	96.06 ± 0.78	14.04 ± 5.63
			msu_mfsd	4.76 ± 0.82	98.25 ± 0.27	3.35 ± 0.20
			oulu_npu	9.07 ± 1.99	96.35 ± 2.21	16.54 ± 5.84
			replay_attack	15.30 ± 0.36	89.33 ± 7.80	5.66 ± 1.15
	0.5	2	casia_fasd	10.19 ± 0.63	96.00 ± 0.13	11.01 ± 0.22
			msu_mfsd	4.84 ± 0.77	98.04 ± 0.85	3.98 ± 1.49
			oulu_npu	9.04 ± 0.70	96.54 ± 0.68	18.19 ± 3.38
			replay_attack	16.43 ± 1.01	89.04 ± 2.78	5.24 ± 0.72
		3	casia_fasd	7.52 ± 0.96	97.23 ± 0.74	8.69 ± 2.14
			msu_mfsd	5.71 ± 1.43	98.08 ± 0.78	4.25 ± 2.29
			oulu_npu	10.28 ± 0.75	95.91 ± 0.76	15.50 ± 6.78
			replay_attack	14.62 ± 0.83	86.02 ± 2.08	4.07 ± 0.91
	1.0	2	casia_fasd	11.33 ± 1.76	95.14 ± 1.76	12.22 ± 2.46
			msu_mfsd	5.16 ± 0.77	97.85 ± 0.63	5.26 ± 1.72
			oulu_npu	8.64 ± 1.20	97.01 ± 0.73	15.94 ± 2.97
			replay_attack	14.60 ± 0.23	85.50 ± 2.16	6.02 ± 1.76
		3	casia_fasd	9.48 ± 1.20	96.51 ± 0.36	11.73 ± 3.17
			msu_mfsd	4.68 ± 0.69	97.91 ± 0.61	4.38 ± 1.46
			oulu_npu	9.89 ± 0.86	96.51 ± 0.72	13.23 ± 3.64
			replay_attack	16.43 ± 1.73	87.15 ± 3.14	5.10 ± 1.25
0.5	0.2	2	casia_fasd	11.41 ± 2.07	95.01 ± 0.64	15.63 ± 3.03
			msu_mfsd	5.32 ± 0.69	98.05 ± 0.31	3.81 ± 0.87
			oulu_npu	9.24 ± 0.97	96.72 ± 0.66	14.36 ± 4.53
			replay_attack	14.77 ± 1.98	87.13 ± 1.67	4.24 ± 0.71
		3	casia_fasd	8.59 ± 0.61	96.99 ± 0.64	15.81 ± 7.16
			msu_mfsd	4.68 ± 0.90	97.56 ± 0.77	7.23 ± 3.03
			oulu_npu	9.73 ± 2.15	96.35 ± 1.31	17.46 ± 8.56
			replay_attack	14.53 ± 0.87	84.58 ± 6.36	4.35 ± 0.32
	0.5	2	casia_fasd	10.93 ± 1.67	95.85 ± 0.68	13.50 ± 0.68
			msu_mfsd	4.29 ± 0.00	98.36 ± 0.25	4.75 ± 2.12
			oulu_npu	9.56 ± 0.84	96.50 ± 0.55	14.65 ± 5.70
			replay_attack	15.50 ± 2.17	88.40 ± 0.87	4.25 ± 0.96
		3	casia_fasd	9.33 ± 0.11	96.07 ± 0.38	11.58 ± 3.69
			msu_mfsd	4.84 ± 0.77	98.39 ± 0.27	4.17 ± 1.06
			oulu_npu	9.78 ± 1.00	96.40 ± 0.48	15.61 ± 3.30
			replay_attack	15.67 ± 0.58	87.55 ± 4.36	5.13 ± 1.63
	1.0	2	casia_fasd	11.67 ± 0.29	94.96 ± 0.51	13.66 ± 3.43
			msu_mfsd	5.24 ± 0.82	98.64 ± 0.28	3.28 ± 1.19
			oulu_npu	8.55 ± 0.32	97.11 ± 0.12	14.42 ± 2.91
			replay_attack	14.80 ± 1.64	87.56 ± 0.25	3.59 ± 0.69
		3	casia_fasd	7.78 ± 0.78	97.16 ± 0.18	10.48 ± 1.41
			msu_mfsd	4.29 ± 0.00	98.36 ± 0.27	5.15 ± 1.96
			oulu_npu	9.63 ± 2.38	96.57 ± 1.46	12.17 ± 4.32
			replay_attack	14.82 ± 1.30	91.77 ± 0.87	4.90 ± 1.16
1.0	0.2	2	casia_fasd	11.37 ± 0.06	94.82 ± 0.18	13.38 ± 1.98
			msu_mfsd	5.24 ± 1.45	97.81 ± 1.20	3.91 ± 1.45
			oulu_npu	9.27 ± 1.28	96.90 ± 0.63	15.83 ± 3.00
			replay_attack	14.43 ± 1.09	87.53 ± 1.09	5.14 ± 1.57
		3	casia_fasd	9.33 ± 1.33	96.20 ± 1.01	12.95 ± 3.33
			msu_mfsd	4.76 ± 1.45	97.71 ± 0.32	3.91 ± 0.86
			oulu_npu	9.87 ± 1.52	96.22 ± 0.88	15.70 ± 2.43
			replay_attack	16.42 ± 0.53	83.62 ± 3.10	4.36 ± 0.36
	0.5	2	casia_fasd	11.30 ± 1.03	95.26 ± 0.57	13.71 ± 2.89
			msu_mfsd	5.24 ± 0.63	97.98 ± 0.36	4.20 ± 1.24
			oulu_npu	8.72 ± 1.30	97.16 ± 1.00	17.05 ± 5.28
			replay_attack	13.60 ± 0.35	88.61 ± 4.60	5.37 ± 1.18
		3	casia_fasd	8.37 ± 0.84	97.11 ± 0.57	13.56 ± 6.12
			msu_mfsd	4.68 ± 0.73	98.25 ± 0.17	5.09 ± 2.66
			oulu_npu	10.28 ± 2.67	95.86 ± 1.91	14.59 ± 5.60
			replay_attack	15.53 ± 1.33	90.68 ± 4.80	4.83 ± 0.35
	1.0	2	casia_fasd	11.11 ± 0.38	95.47 ± 0.43	12.27 ± 0.91
			msu_mfsd	5.24 ± 0.82	97.55 ± 0.73	3.77 ± 1.02
			oulu_npu	8.78 ± 0.59	97.14 ± 0.36	15.05 ± 3.13
			replay_attack	14.38 ± 0.33	88.06 ± 2.91	4.60 ± 0.55
		3	casia_fasd	8.85 ± 0.23	97.09 ± 0.36	12.39 ± 1.04
			msu_mfsd	4.68 ± 0.90	98.22 ± 0.87	4.98 ± 3.75
			oulu_npu	10.37 ± 3.19	96.12 ± 1.98	11.73 ± 3.35
			replay_attack	14.62 ± 1.81	91.02 ± 1.98	4.61 ± 1.52

Table 1: Mean HTER and AUC (± standard deviation) for various combinations of loss weights and segmentation output channels. The best HTER results for each dataset are highlighted by color (red = CASIA-FASD, blue = Idiap Replay-Attack, green = MSU-MFSD, yellow = OULU-NPU). Results show mean ± std of HTER (%) and AUC (%) over 3 independent trials with different random seeds.