

Self-Supervised Emotion Representation Disentanglement for Speech-Preserving Facial Expression Manipulation

Supplementary Material

Anonymous Author(s)

1 QUANTITATIVE COMPARISONS

In the main manuscript, we present average metrics for comparing our SSERD framework with current state-of-the-art methods. Here, we provide detailed metrics for each emotion to facilitate comprehensive comparisons, as shown in Tables 1 and 2. Our framework demonstrates superior performance across various emotions in all three metrics. For instance, considering the 'Fear' emotion, our framework shows a notable improvement over the NED method in the MEAD dataset. In the intra-identity setting, it reduces FAD and LSE-D by 52.9% and 9.7%, respectively, while increasing CSIM by 5.2%. In the cross-identity setting, it decreases FAD and LSE-D by 55.5% and 3.9%, respectively, and enhances CSIM by 1.9%. Similarly, compared to the DSM method on the RAVDESS dataset, our framework reduces FAD and LSE-D by 43.7%

and 4.2%, respectively, and increases CSIM by 5.2% in the intra-identity setting. In the cross-identity setting, it decreases FAD and LSE-D by 39.3% and 5.7%, respectively, and improves CSIM by 1.3%.

2 QUALITATIVE COMPARISONS

In the main manuscript, we present some visualization examples for qualitative comparisons between current state-of-the-art methods and our SSERD framework. Expanding upon this, we provide additional examples on the MEAD and RAVDESS datasets, as depicted in Figures 1, and 2. Consistent with the findings reported in the main manuscript, our framework demonstrates superior image realism, lip synchronization, and emotion similarity across both datasets.

Settings	Metrics	Methods	Neutral	Angry	Disgusted	Fear	Happy	Sad	Surprised	Avg.
Intra-ID	FAD↓	ICface	7.114	6.420	7.383	6.567	6.213	7.301	5.567	6.795
		DSM	2.572	2.156	2.125	2.364	1.951	1.985	1.908	2.152
		NED	0.906	2.177	3.838	1.659	1.939	2.538	1.700	2.108
		Ours	0.473	0.683	0.820	0.781	0.648	0.808	0.965	0.740
	LSE-D↓	ICface	9.760	10.483	10.433	9.855	10.180	10.017	9.851	10.083
		DSM	9.452	9.835	9.272	9.676	9.664	9.594	9.266	9.531
		NED	9.264	9.579	9.128	10.172	9.137	9.074	9.821	9.454
		Ours	8.964	9.188	8.796	9.181	9.315	9.253	9.188	9.126
	CSIM↑	ICface	0.779	0.741	0.805	0.754	0.775	0.755	0.817	0.775
		DSM	0.806	0.780	0.815	0.790	0.815	0.821	0.818	0.806
		NED	0.883	0.802	0.772	0.848	0.839	0.812	0.864	0.831
		Ours	0.908	0.874	0.929	0.892	0.908	0.914	0.906	0.904
Cross-ID	FAD↓	ICface	10.560	9.470	9.230	9.122	8.493	10.364	9.541	9.540
		EAT	5.354	7.103	6.737	6.198	6.160	6.114	5.635	6.186
		DSM	1.916	5.071	4.991	4.686	5.274	4.943	4.338	4.460
		NED	2.022	4.851	5.094	4.983	3.919	5.665	4.600	4.448
		Ours	0.490	3.308	4.011	2.215	2.307	2.483	2.360	2.453
	LSE-D↓	ICface	11.226	11.073	11.184	11.204	11.322	11.526	11.133	11.238
		EAT	9.616	9.653	9.574	9.435	9.562	9.507	9.508	9.551
		DSM	9.801	9.888	10.157	9.739	9.518	9.961	10.357	9.917
		NED	9.812	9.904	10.121	9.741	9.936	10.179	9.646	9.906
		Ours	8.956	9.016	9.051	9.364	9.304	9.299	9.410	9.200
	CSIM↑	ICface	0.705	0.648	0.637	0.727	0.717	0.664	0.721	0.688
		EAT	0.761	0.736	0.756	0.764	0.791	0.696	0.823	0.761
		DSM	0.866	0.753	0.784	0.737	0.777	0.744	0.787	0.778
		NED	0.841	0.717	0.791	0.750	0.842	0.691	0.780	0.773
		Ours	0.908	0.799	0.843	0.837	0.858	0.838	0.852	0.848

Table 1: Comparison results of FAD, LSE-D and CSIM of our framework and competing methods in the intra-identity and cross-identity settings on the MEAD dataset.

Settings	Metrics	Methods	Neutral	Angry	Disgusted	Fear	Happy	Sad	Surprised	Avg.
Intra-ID	FAD↓	ICface	9.816	7.047	8.689	8.413	8.413	8.086	8.636	8.443
		DSM	1.653	2.833	3.161	2.122	2.186	1.937	2.601	2.354
		NED	2.041	3.288	4.144	2.635	3.714	2.595	2.980	3.057
		Ours	1.021	1.127	1.287	1.483	1.161	1.999	1.718	1.399
	LSE-D↓	ICface	8.209	9.504	8.296	8.523	8.902	8.346	7.578	8.480
		DSM	7.697	7.860	7.698	7.201	7.820	7.656	7.641	7.653
		NED	7.376	7.757	7.822	7.452	7.742	7.560	7.226	7.562
		Ours	7.578	7.786	7.428	7.136	7.472	7.363	7.326	7.441
	CSIM↑	ICface	0.749	0.703	0.775	0.722	0.797	0.766	0.772	0.755
		DSM	0.879	0.812	0.889	0.857	0.895	0.899	0.865	0.871
		NED	0.847	0.805	0.786	0.842	0.793	0.855	0.848	0.825
		Ours	0.898	0.810	0.927	0.886	0.914	0.932	0.894	0.894
Cross-ID	FAD↓	ICface	10.478	8.704	9.260	9.106	9.061	9.639	9.718	9.424
		EAT	6.680	8.384	8.591	8.655	8.239	7.937	7.914	8.051
		DSM	1.883	4.705	6.865	4.802	3.818	4.438	3.292	4.258
		NED	3.558	5.546	5.008	5.648	5.648	5.588	5.145	5.412
		Ours	1.085	4.045	5.785	3.431	3.336	3.063	2.772	3.360
	LSE-D↓	ICface	10.736	12.415	11.860	11.279	11.150	11.305	12.028	11.539
		EAT	8.223	8.028	8.089	8.085	8.369	8.203	8.080	8.154
		DSM	8.141	8.357	8.286	8.247	8.277	8.048	8.109	8.209
		NED	7.856	8.085	8.107	8.151	8.073	8.006	7.962	8.034
		Ours	7.557	7.587	7.595	7.690	7.524	7.693	7.704	7.621
	CSIM↑	ICface	0.677	0.646	0.717	0.649	0.738	0.666	0.644	0.677
		EAT	0.697	0.636	0.674	0.574	0.785	0.632	0.679	0.668
		DSM	0.848	0.707	0.694	0.732	0.801	0.745	0.762	0.756
		NED	0.820	0.766	0.741	0.749	0.804	0.726	0.713	0.760
		Ours	0.898	0.737	0.676	0.759	0.809	0.856	0.794	0.790

Table 2: Comparison results of FAD, LSE-D and CSIM of our framework and competing methods in the intra-identity and cross-identity settings on the RAVDESS dataset.

Emotions	Realism					Emotion similarity					Mouth shape similarity				
	ICface	EAT	DSM	NED	Ours	ICface	EAT	DSM	NED	Ours	ICface	EAT	DSM	NED	Ours
Neutral	4%	14%	11%	7%	64%	4%	11%	11%	25%	50%	4%	14%	14%	14%	54%
Angry	0%	5%	26%	33%	37%	2%	14%	21%	33%	30%	0%	9%	35%	21%	35%
Disgusted	2%	24%	29%	20%	24%	0%	20%	27%	27%	27%	0%	12%	29%	24%	34%
Fear	0%	9%	24%	9%	58%	0%	20%	11%	18%	51%	0%	9%	22%	16%	53%
Happy	0%	14%	26%	21%	40%	0%	23%	21%	16%	40%	0%	7%	30%	19%	44%
Sad	0%	13%	36%	15%	36%	3%	10%	26%	18%	44%	3%	10%	23%	15%	49%
Surprised	2%	17%	29%	24%	27%	2%	34%	24%	22%	17%	2%	20%	17%	24%	37%
All	1%	14%	26%	19%	40%	1%	19%	20%	22%	36%	1%	11%	25%	19%	43%

Table 3: Realism, emotion similarity, and mouth shape similarity ratings of each emotion of the user study.

3 USER STUDY & VIDEO EXAMPLES

In the main manuscript, we present the aggregate voting results from our user study. This section provides a detailed breakdown of the voting results for each emotion. As illustrated in Table 3, our framework achieves higher ratings for mouth shape similarity across various emotions. Additionally, it receives the highest ratings for image realism and emotional similarity in the 'Neutral', 'Fear', 'Happy', and 'Sad' categories. However, our framework exhibits a limitation in accurately expressing the 'Surprise' emotion, which appears inconsistent with the results presented in Tables

1 and 2. This discrepancy likely stems from the variability in the expressiveness of the actors within the dataset. In some cases, the expressions of surprise conveyed by the actors do not seem entirely appropriate, according to our assessment. Therefore, although our method achieves superior quantitative metrics by generating images that more closely align with the real distribution, this fidelity results in lower scores when evaluated by human judges.

All videos used in our user study are included in the supplementary materials for review purposes. On the online platform, we have carefully organized the videos from different methods to

ensure unbiased voting outcomes. For clearer visual comparison, the videos in the supplementary materials are arranged as follows: the first and second columns display the source and reference

images, respectively, while the subsequent columns show the results generated by ICface, DSM, NED, EAT, and our framework.

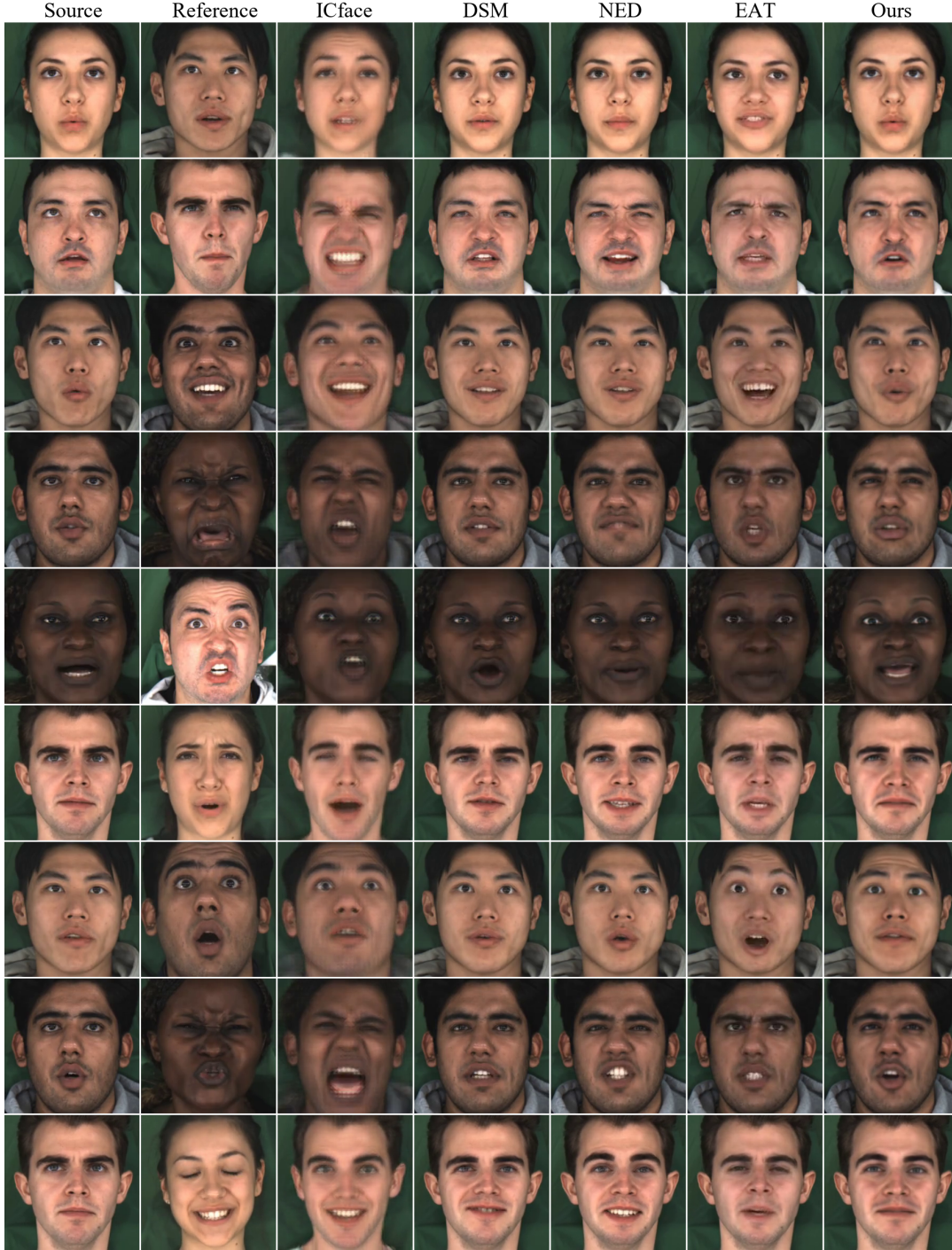


Figure 1: Qualitative comparisons with state-of-the-art methods on the MEAD dataset. Our framework produces expressive talking faces featuring natural expressions and synchronized lip movements.



Figure 2: Qualitative comparisons with state-of-the-art methods on the RAVDESS dataset. Our framework produces expressive talking faces featuring natural expressions and synchronized lip movements.