Discrete Facial Encoding: A Framework for Data-driven Facial Display Discovery (Supplementary materials)

A. Ablation Studies

We conducted an ablation study on the StressID dataset to estimate the impact of model design choices. Results are shown in Table 1. Decreasing the number of quantizers has the largest effect: reducing from four to a single quantizer, effectively turning it into a VQ-VAE, results in a significant drop in performance. Codebook size also influences performance—both overly small (16) and overly large (256) codebooks result in lower performance. We chose a simple decoder, as the focus of this approach is on the encoder. To demonstrate this, we trained the same model with a more complex and deeper decoder (a 6-layer transformer with a hidden dimension of 128 and 4 attention heads). The larger decoder does not result in improved downstream performance, demonstrating the adequacy of a simple decoder in enabling the training of our encoder. Removing the orthogonality or L1 loss leads to improved performance on the binary Stress ID task but reduced performance on the multiclass Stress ID task; in addition, this trade-off is associated with decreased model interpretability.

In Figure 1 we show that these design choices also affect the learned codebook representations. Without orthogonality loss, the regions of interest captured by different codewords largely overlap, leading to redundant templates. Without sparsity loss, the regions cover broad global areas of the face rather than focusing on local discriminative regions, which reduces interpretability. Finally, with only one quantizer, the model fails to capture diverse facial patterns; faces become non-decomposable to a combination of templates, limiting representational capacity to a single global template per face. We also plot the percentile curve representing the distribution of vertex displacements between the learned facial codebook mesh and the neutral mesh (Figure 2), as detailed in Section 3.3. This visualization illustrates the number of vertices that undergo a given amount of displacement, providing insight into the variability of the rendered vertices in the learned facial templates. Notably, our method consistently yields the lowest curve, indicating that it produces significantly fewer vertices with large displacements compared to the other two ablation settings. This result demonstrates that our rendered mesh is substantially less scattered.

In Table 2, we evaluate the impact of various components on the orthogonality of the learned representations. Specifically, we assess the similarity between the displacement vectors of facial templates associated with each codeword. To quantify this, we compute both the dot product and cosine similarity for all pairs of codewords in the codebook, reporting the average values. Higher scores indicate greater similarity (i.e., more redundant templates), whereas lower scores reflect increased diversity among the templates. Our method achieves the lowest average dot product and cosine similarity, showing lower redundancy and more unique facial templates compared to configuration without sparsity loss.

Table 1. Performance comparison on the StressID dataset. We report F1 Score and Balanced Accuracy for binary and multiclass classification. All values are multiplied by 100 for readability.

2*Model	Binary		Multiclass	
	F1 ↑	BAcc ↑	F1 ↑	BAcc ↑
Ours w/ Codebook Size 256	71.4	71.0	56.2	56.0
Ours w/ Codebook Size 16	73.7	73.3	59.6	59.0
Ours w/ Single Quantizer (VQ-VAE)	70.2	69.9	52.3	51.5
Ours w/ Transformer Decoder	73.1	72.8	57.2	56.6
Ours w/o orthogonality loss	76.0	75.7	59.4	58.8
Ours w/o L1 loss	**77.4**	**77.0**	57.9	57.5
Ours	73.8	73.5	**60.3**	**59.7**

Table 2. Displacement regions orthogonality comparison on the StressID dataset. We compute dot product and cosine similarity of face displacement vectors corresponding to different codewords. Lower values indicate higher diversity.

Model	Dot product ↓	Cosine ↓
Single Quantizer (VQ-VAE)	0.0493	0.8676
Ours w/o orthogonality loss	0.0158	0.8468
Ours w/o L1 loss	0.0171	0.8390
Ours	**0.0086**	**0.8268**

B. Visualization of all learned facial templates

We provide a visualization of all learned facial templates in Figure 3. Our system successfully captures high-frequency

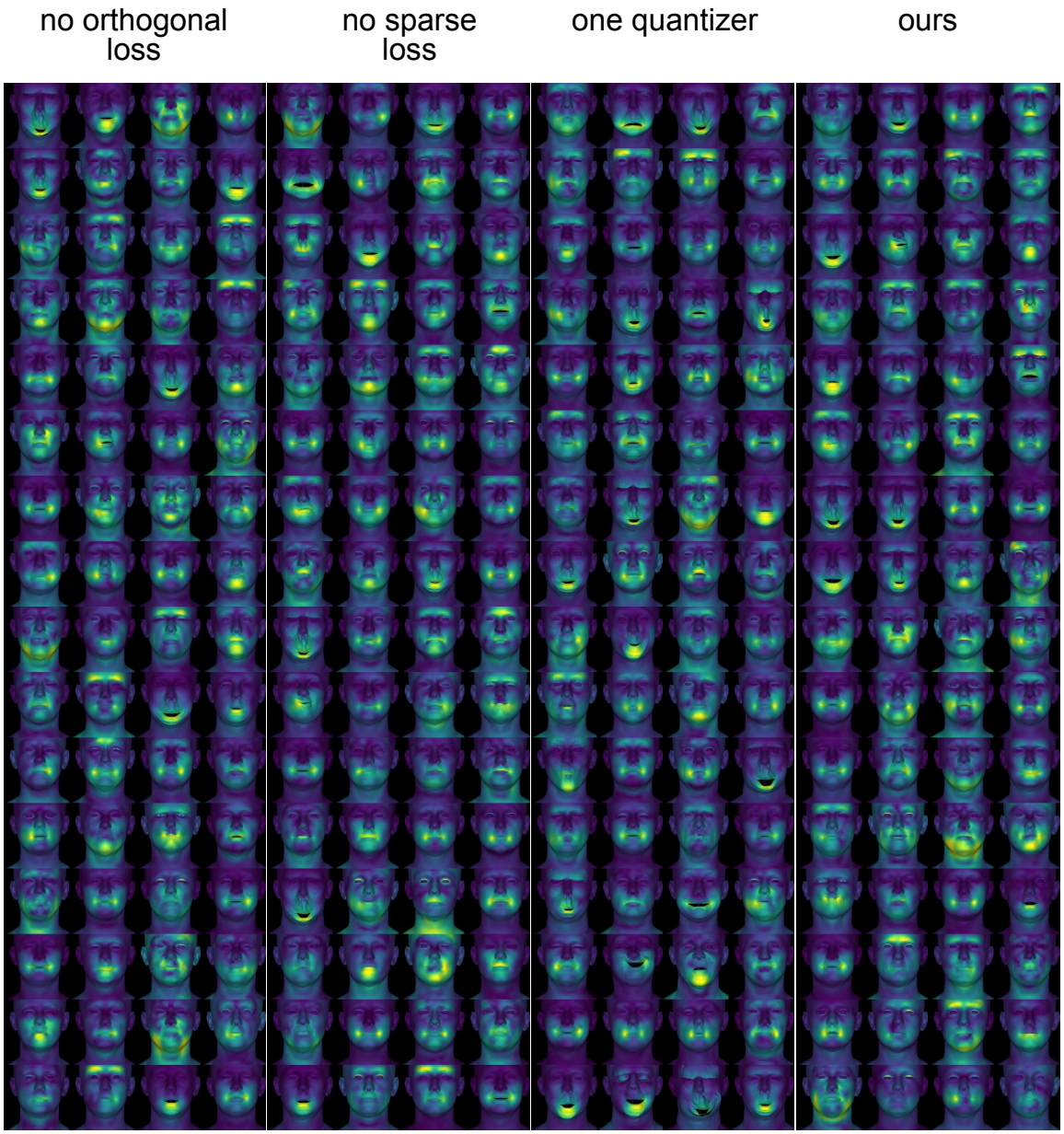

Figure 1. Impact of design choices on learned codewords. Without orthogonality loss, codewords capture overlapping regions, resulting in redundant templates. Without sparsity loss, regions expand to broad facial areas rather than local discriminative features, reducing interpretability. With only one quantizer, the model fails to capture diverse patterns, yielding a single global template per face instead of a compositional representation.

facial movements, including both symmetric and asymmetric motions. This stands in contrast to existing automated Action Unit tools, which are trained on symmetric annotations and thus tend to be biased toward decoding only symmetric facial motions.

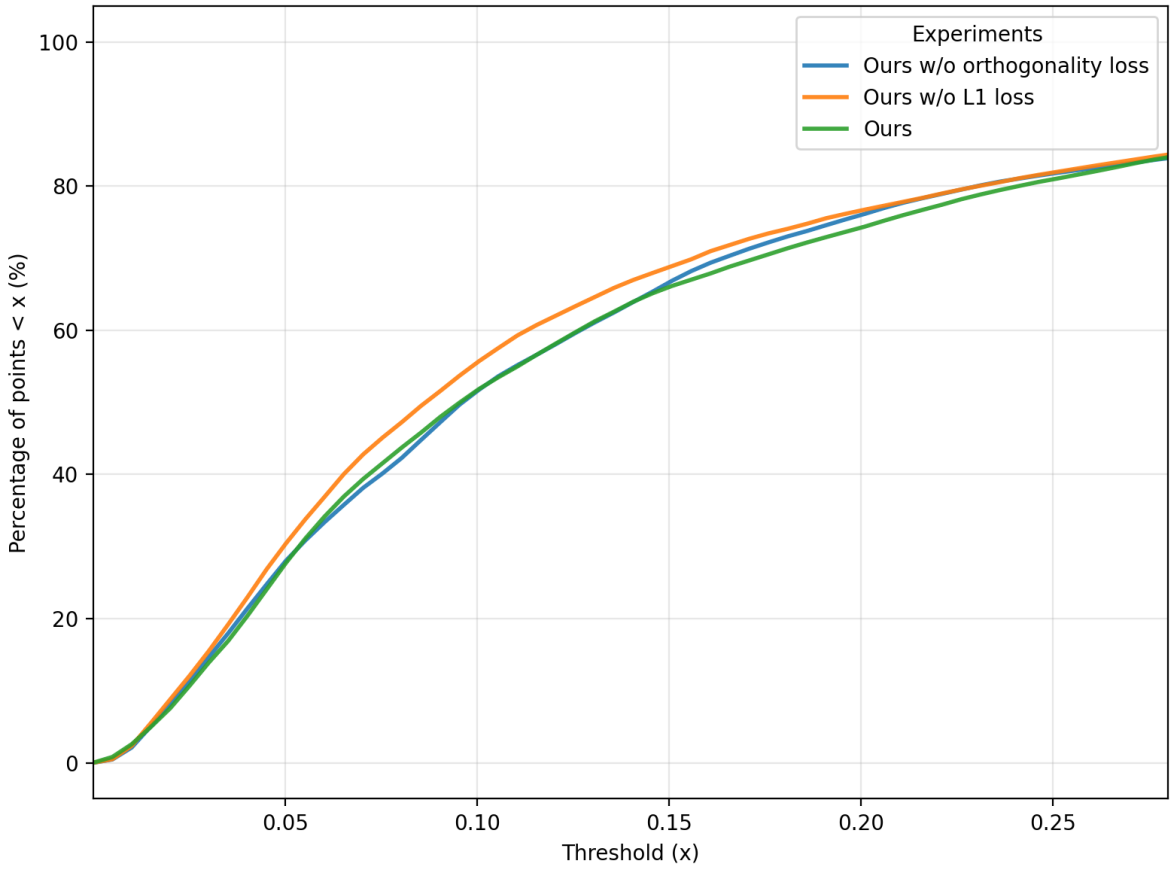

Figure 2. Percentile curve of displacement points distribution. Shows percentage of points with displacements grater than current value

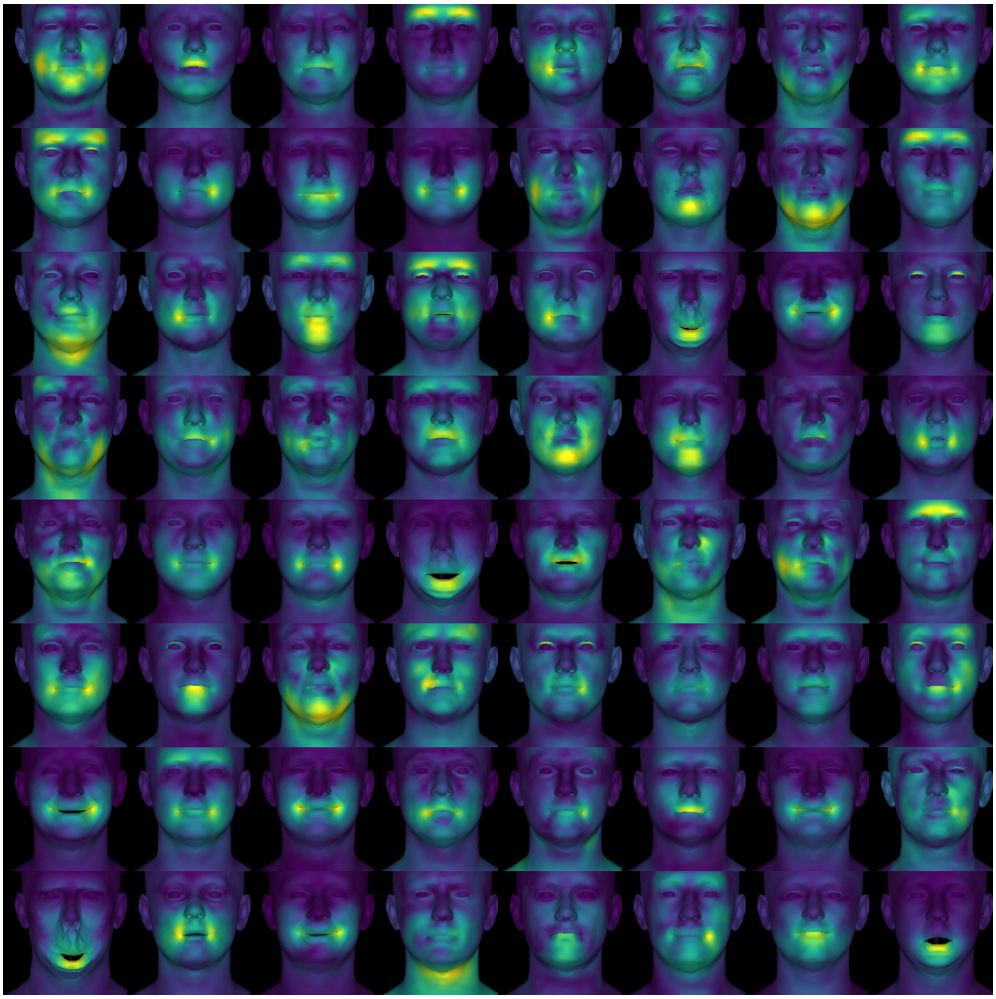

Figure 3. Visualization of the learned facial templates.