BlendME: A VR Blendshape Microexpression Dataset with Retrospective Self-Annotation

Supplementary Material

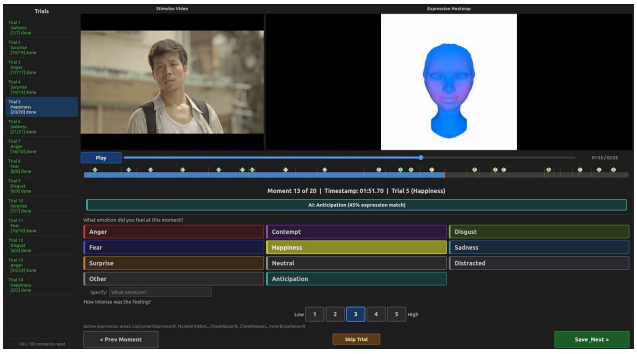

Figure 1. Post-session retrospective annotation interface. (a) Participant reviewing detected key moments at the desktop workstation. (b) Annotation UI displaying stimulus video, expression heatmap, AI-suggested emotion, and intensity rating controls.

1. Retrospective Annotation Interface

Figure 1 shows the post-session retrospective annotation interface. Participants view detected activation peaks on a playback timeline alongside the stimulus video and an avatar replaying their recorded blendshape activations. For each peak, they assign an emotion label from 11 categories and an intensity rating.

2. Stimulus Validation

Figure 2 shows hit rates, mean emotion intensity, and suppression difficulty for all 37 candidate videos tested across three pilot rounds ($N = 15$ each). The base design used two videos per emotion (12 total). Because anger and surprise consistently showed the lowest pilot hit rates (47% and 67% respectively), an additional backup video was included for each, bringing the final set to 14 stimuli. Happiness and disgust stimuli consistently exceeded the 70% threshold, anger remained the most difficult to elicit, with no candidate surpassing 47%.

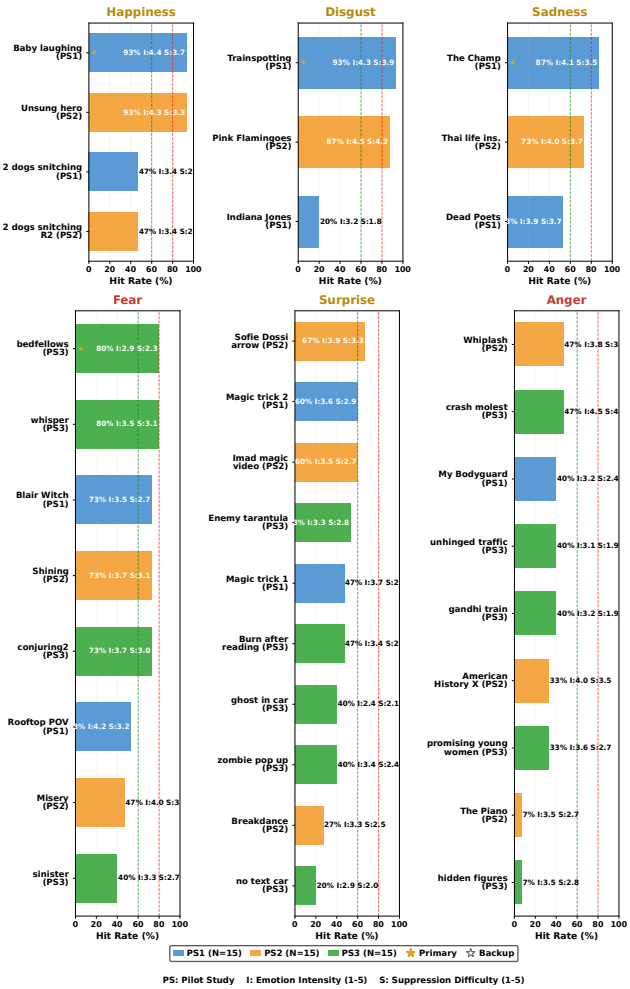

Figure 2. Video stimuli validation across three pilot studies (N=15 each). Hit rate, mean emotion intensity (I), and suppression difficulty (S) for 37 candidate videos across six target emotions. Dashed lines indicate 50% (red) and 70% (green) hit rate thresholds. Stars denote primary and backup selections for the final 14-video set

3. Blendshape ME Schematic

Figure 3 illustrates a single microexpression as it appears in the blendshape time series. The top-5 activated channels from the 70-dimensional blendshape vector are shown over approximately 1200 ms. The event spans 333 ms (15 frames at 45 Hz) from onset to offset, with a clear asymmetric profile: rapid onset reaching apex within the first half of the event, followed by a slower return to baseline. This temporal shape is characteristic of microexpressions and is the signal that TRM's bandpass filter and energy computation (Section 4.2) are designed to detect.

4. ME Detection Pipeline Details

Quest Pro microexpressions are detected using a FACS-guided pipeline operating on TRM peaks. Each candidate must satisfy strict ME criteria: duration 3–25 frames (67–556 ms at ∼45 Hz), peak amplitude between $1.5 \times$SD above

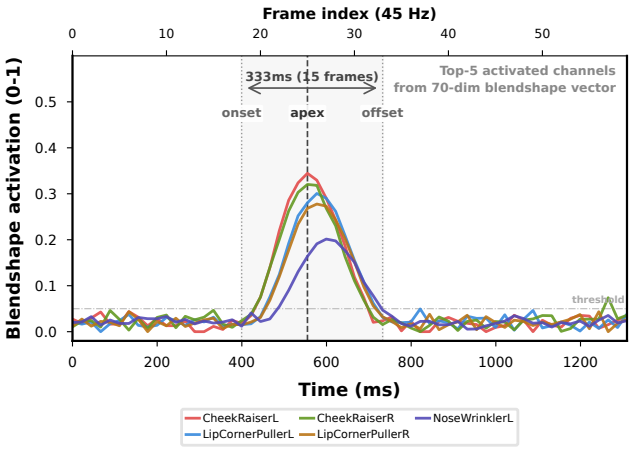

Figure 3. Schematic of a single microexpression in 1D blendshape time series format. Top-5 activated channels from the 70-dimensional blendshape vector are plotted. The event spans 333 ms (15 frames at 45 Hz) with onset, apex, and offset marked. The asymmetric temporal profile (rapid onset, slower offset) is characteristic of MEs.

neutral baseline and < 0.4 (macro threshold), unimodal temporal profile, 1–6 coherent directly-observed channels, and baseline return within 1 SD in 3 frames post-offset. A FACS gate ensures each candidate's dominant blendshape activation matches the canonical AU pattern for its labeled emotion. Velocity-based blink rejection uses inter-frame eyelid displacement rather than absolute thresholds. Repetitive jaw/brow movement patterns ($\geq$5 similar peaks in 10 s) are rejected unless they contain emotion-specific channels. Emotion labels derive from video event annotations within a 3 s reaction window, following the stimulus-based labeling convention used by CASME II and SAMM. Where participant self-annotations were available, 73% agreed with the video event label on a basic emotion; the remaining 27% used non-basic categories (e.g., "neutral," "anticipation"), consistent with established difficulty of ME self-labeling.

The same five participants used for TRM parameter tuning (Section 4.3) served as the held-out calibration set for the quality filter grid search. A search over 864 parameter configurations (amplitude threshold, isolation minimum, blink velocity cutoff, baseline return ratio) optimized a multi-objective function combining d' (emotion separability), Spearman ρ (rank agreement with established datasets), emotion balance, and participant diversity on this calibration set. The winning configuration was then applied to all remaining participants, yielding 112 clips from 21 participants. Table 1 shows the full emotion distribution.

Table 1. Emotion distribution of all retrospective-annotated peaks before quality filtering (304 total, 29 participants, 11 categories). The final filtered dataset (Table 3, main paper) contains 112 clips across six basic emotions.

Emotion	N	%	Emotion	N	%
Anger	73	24.0	Surprise	13	4.3
Sadness	59	19.4	Neutral	10	3.3
Happiness	54	17.8	Contempt	4	1.3
Anticipation	37	12.2	Other	3	1.0
Disgust	25	8.2	Distracted	1	0.3
Fear	25	8.2			

5. Cross-Dataset Comparison Details

5.1. Representation Spaces

We compare Quest Pro and reference dataset MEs in three representation spaces:

- **OVR blendshape space** (63D): Native for Quest Pro; FLAME→OVR translated for reference datasets ($R^2 = 34.6\%$). Eye gaze and arousal channels excluded.
- **FLAME expression space** (100D): Native for reference datasets; OVR→FLAME reverse-translated for Quest Pro ($R^2 = 38.9\%$).
- **Vertex space** (15,069D): Both sources projected to FLAME mesh vertex deltas. No information loss for either source.

5.2. Cross-Dataset Metrics

Table 2 shows per-emotion cross-dataset metrics across all three spaces. Vertex space yields the highest cross-clip cosine similarities (mean 0.34), confirming it as the most faithful comparison space.

Table 2. Per-emotion cross-dataset comparison across representation spaces. "Cross" = mean best-match cosine between QP and DS clips of the same emotion.

	OVR (63D)		FLAME (100D)		Vertex
Emotion	Corr	Cross	Corr	Cross	Cross
Happiness	+0.66	0.48	−0.16	0.19	0.48
Surprise	+0.18	0.46	−0.18	0.18	0.26
Anger	+0.16	0.28	−0.14	0.17	0.44
Disgust	−0.02	0.42	−0.06	0.18	0.33
Sadness	−0.13	0.48	−0.15	0.20	0.34
Fear	−0.00	0.07	−0.08	0.18	0.48
Avg	+0.14	0.37	−0.13	0.18	**0.39**

5.3. FLAME-to-OVR AU Retention

The translation retains only 27% of CASME II's annotated AUs and 17.7% of SAMM's, confirming substantial information loss in the blendshape projection.

Table 3. Per-AU retention rate across CASME II and SAMM combined. Retention = proportion of clips where the annotated AU is preserved after FLAME-to-OVR translation.

AU	Name	Annotated	Retained	Rate
12	Lip Corner Puller	53	29	54.7%
1	Inner Brow Raiser	32	16	50.0%
5	Upper Lid Raiser	11	4	36.4%
2	Outer Brow Raiser	39	13	33.3%
10	Upper Lip Raiser	20	6	30.0%
26	Jaw Drop	7	2	28.6%
4	Brow Lowerer	82	18	22.0%
20	Lip Stretcher	8	1	12.5%
7	Lid Tightener	74	7	9.5%
9	Nose Wrinkler	18	1	5.6%
6	Cheek Raiser	13	0	0.0%
14	Dimpler	17	0	0.0%
15	Lip Corner Depressor	8	0	0.0%
17	Chin Raiser	8	0	0.0%
24	Lip Pressor	8	0	0.0%

Table 3 shows per-AU retention rates across both datasets combined. AU12 (Lip Corner Puller) survives best at 54.7%, followed by AU1 (Inner Brow Raiser) at 50.0%. AU6 (Cheek Raiser), AU14 (Dimpler), AU15 (Lip Corner Depressor), AU17 (Chin Raiser), and AU24 (Lip Pressor) are completely lost, falling in the null space of the FLAME-to-OVR projection.

5.4. FACS-to-OVR Channel Mapping

Table 4 maps FACS Action Units to their corresponding OVR blendshape channels and reports the Quest Pro observation method for each. Channels labeled "direct" are tracked by infrared sensors with direct line of sight to the relevant facial region. "Periocular" channels are partially occluded by the headset and rely on sensors positioned around the eye opening. "ML-inferred" channels (brow region) are not directly observed; their values are estimated from surrounding sensor data using Meta's internal machine learning model.

5.5. Within-Dataset Emotion Discrimination

The lower d' for Quest Pro reflects that headset-captured microexpressions are inherently more subtle than video-tracked expressions, and the ensemble detector's arousal-based selection produces clips with a shared baseline activation pattern. After removing this shared component via chi-square residual analysis (Section 5), clear emotion-specific patterns emerge.

5.6. Translation Pipeline

Reference datasets are processed through a three-stage pipeline:

Table 4. FACS Action Unit to OVR blendshape mapping with Quest Pro observation method.

AU	Name	OVR Channels	Method
1	Inner Brow Raiser	InnerBrowRaiser L/R	ML-inferred
2	Outer Brow Raiser	OuterBrowRaiser L/R	ML-inferred
4	Brow Lowerer	BrowLowerer L/R	ML-inferred
5	Upper Lid Raiser	UpperLidRaiser L/R	Periocular
6	Cheek Raiser	CheekRaiser L/R	Periocular
7	Lid Tightener	LidTightener L/R	Periocular
9	Nose Wrinkler	NoseWrinkler L/R	Direct
10	Upper Lip Raiser	UpperLipRaiser L/R	Direct
12	Lip Corner Puller	LipCornerPuller L/R	Direct
14	Dimpler	Dimpler L/R	Direct
15	Lip Corner Depressor	LipCornerDepressor L/R	Direct
17	Chin Raiser	ChinRaiser T/B	Direct
20	Lip Stretcher	LipStretcher L/R	Direct
23	Lip Tightener	LipTightener L/R	Direct
24	Lip Pressor	LipPressor L/R	Direct
26	Jaw Drop	JawDrop	Direct
28	Lip Suck	LipSuck LB/RB/LT/RT	Direct

Table 5. Within-dataset emotion separability (d') under two channel configurations. "Discr." excludes BrowLowerer and eye gaze channels, which carry shared headset-pressure activation across all emotions. Removing these substantially increases d', with Quest Pro exceeding camera-based benchmarks on discriminative channels.

Dataset	Channels	Within	Between	d'
Quest Pro	All 63†	0.762	0.710	0.18
Quest Pro	Discr.	0.426	0.051	0.96
CASME II/SAMM/SMIC	Discr.	0.361	0.136	0.72

Within = within-emotion cosine similarity. Between = between-emotion cosine similarity. Higher d' = better separation. Discr. = discriminative (BrowLowerer and eye gaze excluded). †All-channel values computed on pre-filter candidates to illustrate the general BrowLowerer hardware effect.

Table 6. Temporal characteristics of Quest Pro ME clips (112 clips, 21 participants).

Emotion	Duration (ms)	N frames	N clips
Anger	144 ± 84	11 ± 5	3
Disgust	163 ± 71	12 ± 4	18
Fear	105 ± 52	8 ± 3	7
Happiness	155 ± 84	11 ± 5	36
Sadness	116 ± 74	8 ± 4	42
Surprise	117 ± 46	9 ± 4	6
All	$\mathbf{136 \pm 77}$	$\mathbf{10 \pm 5}$	**112**

1. **MICA**: Identity extraction from video frames ($\rightarrow$ 300D shape code)
2. **Metrical Tracker**: Per-frame FLAME fitting ($\rightarrow$ 100 expression + 3 jaw + 6 eye parameters)
3. **FLAME$\rightarrow$OVR Translation**: Regularized L-BFGS-B solve projecting FLAME expression deltas onto the 70-channel OVR blendshape basis ($R^2 = 34.6\%$, RMSE $= 0.69$ mm)

For Quest Pro$\rightarrow$FLAME reverse translation: pre-

Table 7. Meta Quest Pro OVR Face Tracking: 70 Blendshape Channels (XrFaceExpression2FB). Channels 0–62 encode facial movement (63 used in this work); channels 63–69 encode tongue position.

#	Blendshape	#	Blendshape
0	BrowLowererL	35	LipFunnelerLT
1	BrowLowererR	36	LipFunnelerRB
2	CheekPuffL	37	LipFunnelerRT
3	CheekPuffR	38	LipPressorL
4	CheekRaiserL	39	LipPressorR
5	CheekRaiserR	40	LipPuckerL
6	CheekSuckL	41	LipPuckerR
7	CheekSuckR	42	LipStretcherL
8	ChinRaiserB	43	LipStretcherR
9	ChinRaiserT	44	LipSuckLB
10	DimplerL	45	LipSuckLT
11	DimplerR	46	LipSuckRB
12	EyesClosedL	47	LipSuckRT
13	EyesClosedR	48	LipTightenerL
14	EyesLookDownL	49	LipTightenerR
15	EyesLookDownR	50	LipsToward
16	EyesLookLeftL	51	LowerLipDepressorL
17	EyesLookLeftR	52	LowerLipDepressorR
18	EyesLookRightL	53	MouthLeft
19	EyesLookRightR	54	MouthRight
20	EyesLookUpL	55	NoseWrinklerL
21	EyesLookUpR	56	NoseWrinklerR
22	InnerBrowRaiserL	57	OuterBrowRaiserL
23	InnerBrowRaiserR	58	OuterBrowRaiserR
24	JawDrop	59	UpperLidRaiserL
25	JawSidewaysLeft	60	UpperLidRaiserR
26	JawSidewaysRight	61	UpperLipRaiserL
27	JawThrust	62	UpperLipRaiserR
28	LidTightenerL	63	TongueTipInterdental
29	LidTightenerR	64	TongueTipAlveolar
30	LipCornerDepressorL	65	TongueFrontDorsalPalate
31	LipCornerDepressorR	66	TongueMidDorsalPalate
32	LipCornerPullerL	67	TongueBackDorsalVelar
33	LipCornerPullerR	68	TongueOut
34	LipFunnelerLB	69	TongueRetreat

Table 8. FLAME Expression Space: 100 PCA Components [1].

Description
FLAME parameterizes facial expression as a 100-dimensional vector $\boldsymbol{\psi} \in \mathbb{R}^{100}$, learned via PCA from $\sim$33,000 registered 4D face scans. Components are ordered by decreasing variance explained. Unlike OVR blendshapes, FLAME expression coefficients are **not** semantically named; each dimension is an orthogonal basis vector mixing multiple facial regions simultaneously. The first $\sim$10 components capture gross jaw opening, smile, and brow movements; higher components encode progressively finer deformations. In practice, most works use the first 50 components ($\boldsymbol{\psi}_{1:50}$), as components 51–100 contribute minimal variance. FLAME additionally parameterizes jaw articulation via 3 rotation parameters separate from the expression vector.

composed augmented basis with vertex-weighted pseudoinverse, per-participant calibration baseline subtraction, tanh soft-clipping at ± 5 ($R^2 = 38.9\%$).

6. Quest Pro Upper vs. Lower Face Tracking

The Meta Quest Pro uses five internal infrared (IR) cameras for face tracking: three directed at the upper face (eyes, brow region) and two at the lower face (mouth, chin). A pre-trained machine learning model converts raw IR images into the 70 blendshape activations listed in Table 7.

The upper and lower face are treated as separate estimation regions with independent confidence scores. The OpenXR specification (XrFaceConfidence2FB) defines two confidence values: `UPPER_FACE` (everything above the upper lip, including eyes and eyebrows) and `LOWER_FACE` (everything below the eyes, including mouth and chin), with the cheek and nose regions overlapping both.

Critically, the brow region (channels 0–1: BrowLowerer, 22–23: InnerBrowRaiser, 57–58: OuterBrowRaiser) is not directly visible to the IR cameras due to occlusion by the headset housing. These channels are ML-inferred: their values are estimated from surrounding periocular sensor data rather than measured by direct optical observation. Meta has not published the architecture or training data of this internal model. Richard et al. [2] independently confirm that HMD cameras cannot capture "broader upper face expressions such as movements in eyebrow, forehead, nose" and treat this as a fundamental hardware limitation of current headset designs.

This distinction is relevant to interpreting our results. The BrowLowerer dominance reported in Section 7.2 and the universal AU6 (Cheek Raiser) inflation reported in Section 6.4 may partly reflect estimation artifacts in the ML-inferred and periocular channels respectively, rather than genuine facial movement. The "direct" lower-face channels (Table 4) are expected to be more reliable, as they are tracked by IR sensors with unobstructed line of sight.

References

[1] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans (FLAME). *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, 36(6), 2017. 4

[2] Alexander et al. Richard. Universal facial encoding of codec avatars from VR headsets. In *ACM SIGGRAPH*, 2024. arXiv:2407.13038. 4