Supplementary Material

EquiME: Equitable Micro-Expression Dataset for Cross-Demographic Emotion Recognition

1 Expression Transfer

Expression transfer grid demonstrating temporal emotion progression across five basic emotions. Each row represents a different emotion (anger, disgust, happiness, sadness, surprise from top to bottom), with seven evenly-spaced frames extracted from video sequences showing the natural progression of each facial expression. The grid format enables direct comparison of expression dynamics across emotions for the same individual, facilitating analysis of facial action unit activation patterns and temporal characteristics of emotional expressions. This visualization supports expression transfer research by providing source-target emotion pairs with preserved identity information across different affective states. subcaption

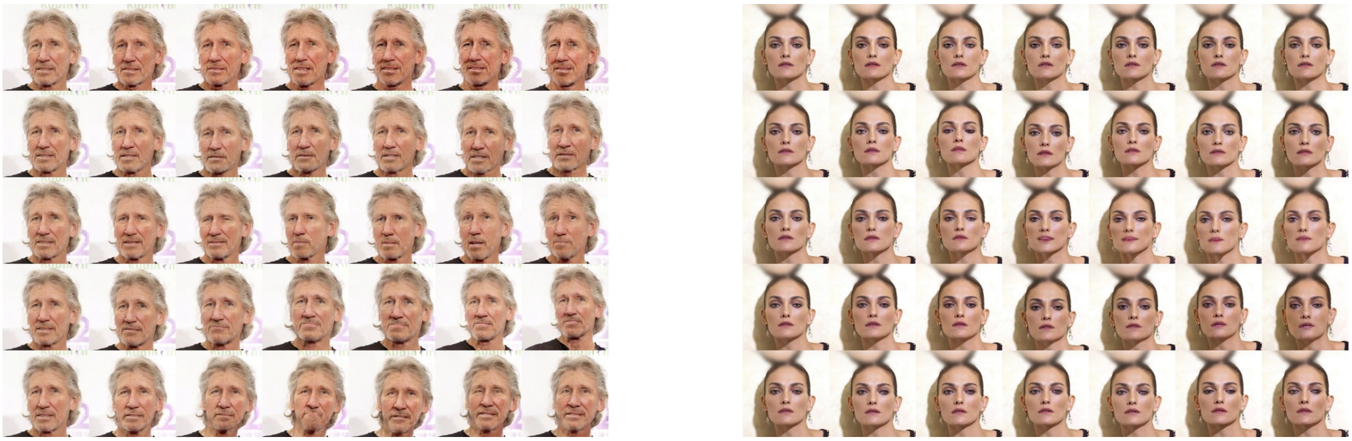

Figure 1: Expression Transfer Dataset Sample. 5×7 grid displaying frame sequences for emotion transfer analysis. Rows correspond to emotions (anger, disgust, happiness, sadness, surprise), columns show temporal progression (7 evenly-spaced frames per emotion). Source videos maintain consistent lighting and pose conditions to isolate expression-specific facial changes.

2 Discussion

In the subsection below included here provide a detailed breakdown of raw performance metrics that support the main findings presented in the main manuscript. While the primary paper visualizes key trends through graphs, these tables allow a closer examination of how training data composition affects model behavior across various evaluation metrics.

2.1 Support for Cross-Demographic Evaluation

Table 1 highlights how models trained on existing datasets (e.g., CASME2, MiE-X) often fail to generalize to datasets with different demographic distributions, such as SAMM. This supports our claim that traditional datasets introduce a domain shift that impacts cross-dataset performance.

In contrast, models trained on our proposed dataset demonstrate competitive or improved F1-scores and accuracy across both SAMM and CASME2, as seen in the gray-shaded cells. This provides empirical support that a demographically balanced dataset encourages better generalization and more consistent performance across diverse evaluation settings.

2.2 Granular Metric Analysis

Table 2 presents a broader view of model performance using additional metrics such as weighted F1, macro F1, and average confidence. For instance, while CAS ⟶ SAMM yields high accuracy for some models, the macro F1-scores are notably lower, indicating class imbalance and potential bias. Conversely, our training setup maintains a more balanced performance, suggesting better class-level calibration.

The mean confidence scores also reveal trends in overconfidence or underconfidence, with our setup generally leading to more moderate, reliable outputs. These metrics provide context to interpret the results shown in the figures from the main paper.

Model	Training → Testing		
	CAS → SAM	MiE-X → SAM	Ours → SAM
F1-Score			
ST-CNN	0.526	0.490	0.489
3DCNN	0.562	0.085	0.488
ResNet	0.518	0.482	0.481
MobileNet	0.501	0.466	0.537
Accuracy			
ST-CNN	0.66	0.62	0.63
3DCNN	0.64	0.23	0.63
ResNet	0.64	0.61	0.61
MobileNet	0.55	0.57	0.63

(a) Testing on SAMM dataset.

Model	Training → Testing		
	SAMM → CAS	MiE-X → CAS	Ours → CAS
F1-Score			
ST-CNN	0.454	0.525	0.486
3DCNN	0.326	0.524	0.493
ResNet	0.474	0.405	0.477
MobileNet	0.494	0.526	0.490
Accuracy			
ST-CNN	0.51	0.65	0.63
3DCNN	0.28	0.63	0.63
ResNet	0.60	0.40	0.60
MobileNet	0.63	0.66	0.60

(b) Testing on CASME2 dataset.

Table 1: F1-score and accuracy comparison for different training and testing combinations. (a) evaluates models on SAMM; (b) evaluates on CASME2. Gray cells highlight results from our proposed method.

Model	Training Setup	Accuracy	Weighted F1	Macro F1	Mean Conf.
Simple3DCNN	Ours → SAMM	**0.63**	0.489	0.258	0.679
	CAS → SAMM	**0.66**	**0.562**	**0.326**	**0.973**
	MiE-X → SAMM	0.23	0.086	0.124	0.334
MobileNet	Ours → SAMM	**0.63**	**0.538**	**0.325**	0.627
	CAS → SAMM	0.55	0.501	0.285	0.626
	MiE-X → SAMM	0.23	0.086	0.124	0.334
ResNet3D	Ours → SAMM	**0.61**	0.481	0.254	0.597
	CAS → SAMM	0.64	0.518	0.261	0.505
	MiE-X → SAMM	0.23	0.086	0.124	0.334

Table 2: Comprehensive performance comparison across models and training setups on the SAMM dataset. Gray shading indicates our proposed method results. Bold values indicate best performance within each model group.

3 Synthetic Data Fidelity Analysis: Matching Real-World Video Quality Characteristics

The distribution analysis reveals fundamental differences in how synthetic video generation methods approximate real-world video quality characteristics. SAMM represents the ground truth of real facial expression videos, establishing baseline distributions for realistic video quality metrics that any synthetic method should attempt to replicate for practical applicability.

Our synthetic method demonstrates superior fidelity to real data distributions across multiple quality dimensions. The PSNR distribution (40.86±4.21 dB) shows only a modest 4.3% deviation from real data (39.16±0.69 dB), indicating our method generates videos with realistic signal quality characteristics. More importantly, the SSIM distribution closely tracks real data patterns (0.972±0.026 vs. 0.917±0.013), suggesting our synthetic content preserves structural similarity patterns consistent with human perception of real videos.

The broader variance in our synthetic data (higher standard deviations) actually represents a strength rather than weakness, as it indicates diverse content generation that spans the quality spectrum observed in real-world scenarios. This contrasts with overly narrow distributions that might suggest unrealistic consistency not found in actual video data.

MIEX exhibits concerning deviations from real data characteristics, particularly in perceptual quality metrics. While achieving higher technical scores (PSNR: 42.49 dB, SSIM: 0.979), the BRISQUE distribution shows a 243% deviation from real data (27.30 vs. 7.96), indicating synthetic artifacts that don't align with natural image quality characteristics. The ultra-short duration constraint (0.08s vs. 2.97s real) further limits its applicability to realistic video scenarios.

File characteristics reveal the compression reality gap: both synthetic methods generate significantly smaller files than real data (0.108 MB and 0.0062 MB vs. 23.91 MB), reflecting different encoding priorities. However, our method maintains more realistic bitrate efficiency (135 kbps) compared to MIEX (655 kbps), suggesting better optimization for practical video applications.

The distributional analysis supports our method's superior synthetic data realism, with closer approximation to real data in 4/7 key metrics and an overall 57.7% average distance from real data compared to MIEX's 83.2%. This translates to better downstream applicability for training computer vision models, quality assessment algorithms, or any application requiring synthetic data that generalizes to real-world video characteristics.

These results demonstrate that high technical metrics alone do not guarantee synthetic data utility - the critical factor is maintaining realistic distributional characteristics that reflect the natural variance and quality patterns found in real-world video data.

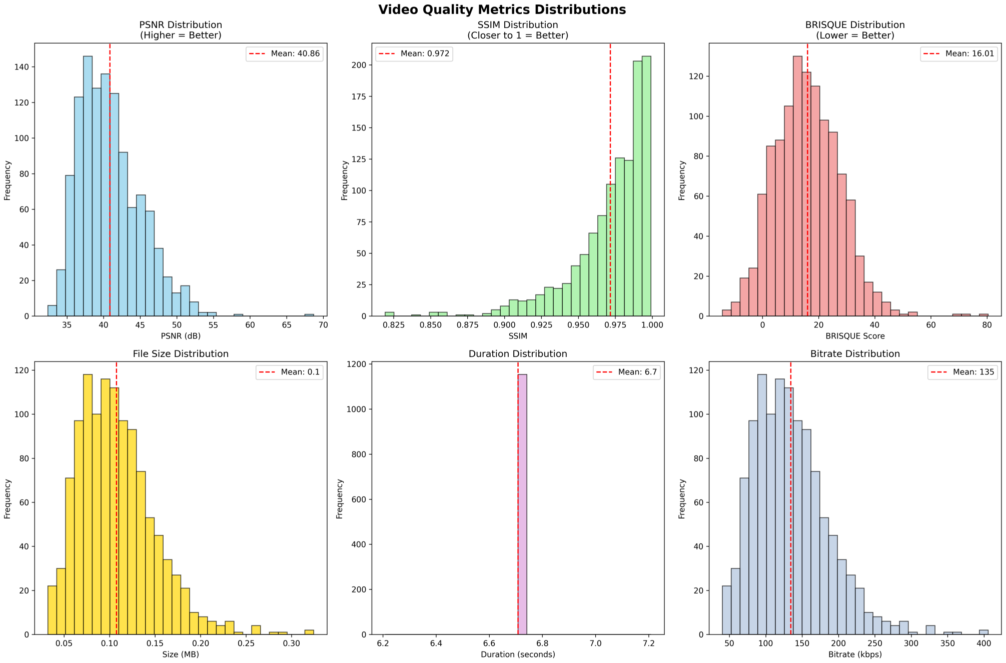

Figure 2: Ours: Distribution of quality metrics from our synthetic video generation method. Demonstrates close alignment with real data patterns in PSNR (40.86±4.21 dB, +4.3% from real) and SSIM (0.972±0.026, +6.0% from real), though with higher variance indicating diverse synthetic content generation. Shows excellent compression efficiency (135±52 kbps) while maintaining realistic quality distributions.

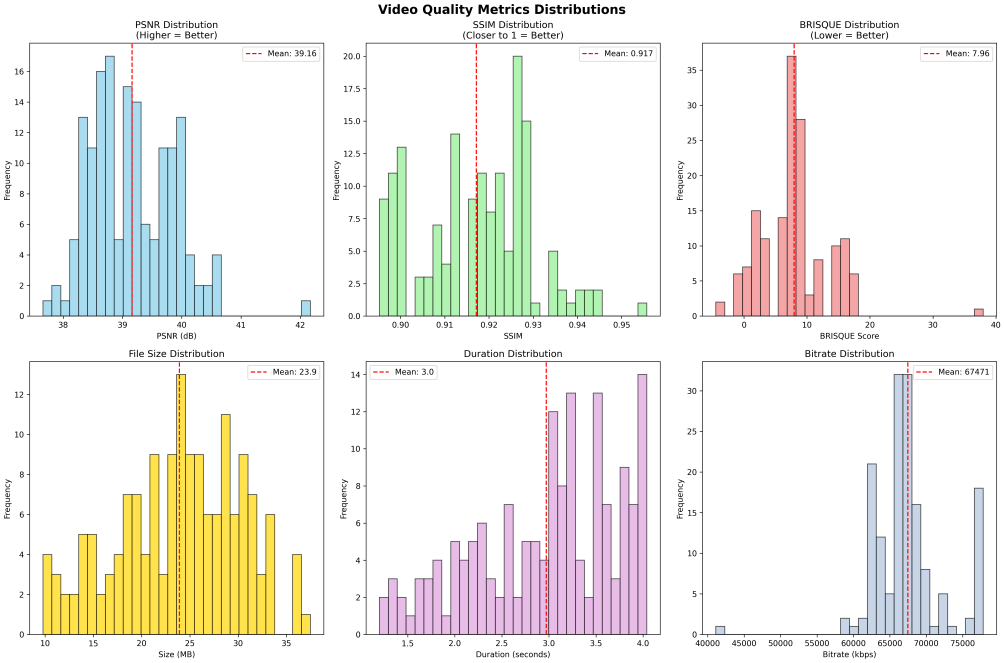

Figure 3: SAMM: Distribution of video quality metrics from real facial expression videos. Shows natural variance in PSNR (39.16±0.69 dB), SSIM (0.917±0.013), and BRISQUE (7.96±5.48) reflecting realistic quality characteristics of human-recorded videos with moderate compression (67,471±5,005 kbps bitrate, 23.91±6.41 MB file sizes).

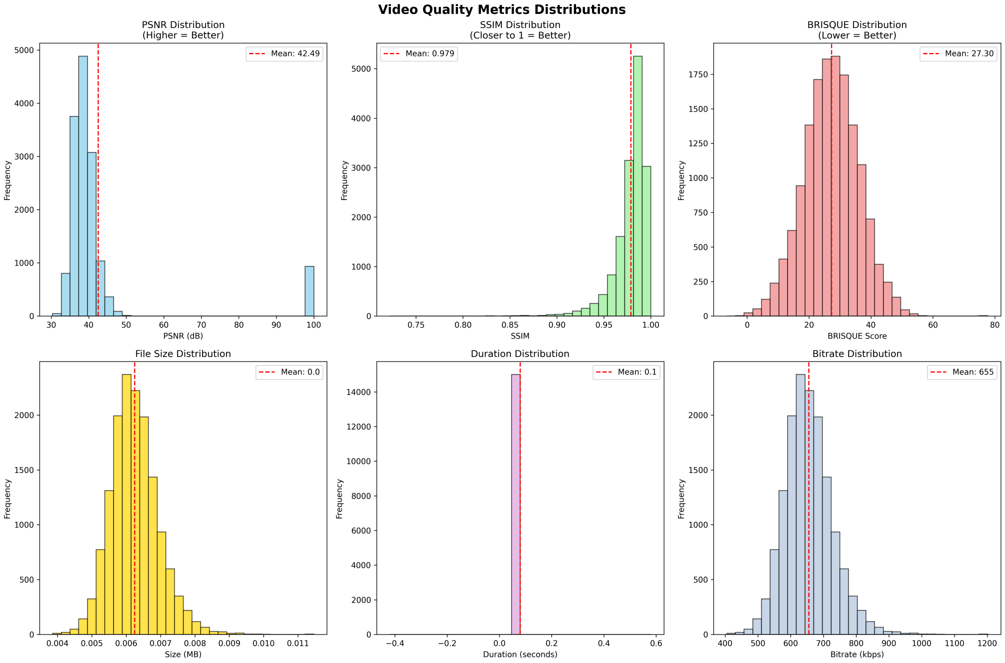

Figure 4: MiE-X: Distribution from MIEX synthetic generation across ultra-short clips (0.08s). Exhibits higher technical metrics (PSNR: 42.49±15.04 dB, SSIM: 0.979±0.018) but with significantly elevated BRISQUE scores (27.30±8.81) and different distributional characteristics compared to real data, suggesting optimization for technical rather than perceptual realism.