rPPG-VQA: A Video Quality Assessment Framework for Unsupervised rPPG Training

Supplementary Material

6. RANSAC Algorithm

The random sample consensus (RANSAC) algorithm is an iterative method for robustly estimating the parameters of a mathematical model from data containing outliers [3]. Its primary strength is its ability to derive a reliable model even when a significant fraction of the dataset consists of erroneous measurements. In this work, we employ RANSAC to determine a single, robust consensus frequency $f_{i,cons}^{HR}$, from a set of M individual estimates $\left\{f_{i,j}^{HR}\right\}_{j=1}^{M}$.

RANSAC operates by assuming the dataset contains both inliers (data that can be explained by the model) and outliers (data that do not fit the model). The algorithm iteratively generates model hypotheses from random data subsets and evaluates them based on the support they receive from the full dataset. We demonstrate the overall algorithm for RANSAC in Algorithm 1.

Algorithm 1 RANSAC Algorithm

Input:
$\mathcal{F} = \left\{f_{i,1}^{HR}, ..., f_{i,M}^{HR}\right\}$: HR frequency estimates
ϵ: inlier tolerance threshold
K: number of iterations
Output: The estimated consensus HR frequency $f_{i,cons}^{HR}$

$s \leftarrow \min(2, M)$
$I^* \leftarrow \varnothing$
for $k = 1$ to K **do**
  $J \subseteq \{1, ..., M\}$ where $|J| = s$
  $\hat{f}_i^{HR} \leftarrow \frac{1}{s} \sum_{j \in J} f_{i,j}^{HR}$
  $I_k \leftarrow \left\{j \in \{1, ..., M\} \mid \left|f_{i,j}^{HR} - \hat{f}_i^{HR}\right| \leq \epsilon\right\}$
  if $|I_k| > |I^*|$ **then**
    $I^* \leftarrow I_k$
  end if
end for
if $I^* \neq \varnothing$ **then**
  $f_{i,cons}^{HR} \leftarrow \text{median}\left(\left\{f_{i,j}^{HR} | j \in I^*\right\}\right)$
else
  $f_{i,cons}^{HR} \leftarrow \text{median}(\mathcal{F})$
end if
return $f_{i,cons}^{HR}$

7. Scene-Level Noise Perception Prompt

The prompt for Qwen3-VL [1] to assess scene-level quality is given in Figure 3.

Table 8. **Ablation study on the fusion size M.**

M	MAE↓	RMSE↓	R↑
1	0.91	1.30	0.99
3	0.78	1.17	0.99
5	0.57	1.12	**1.00**
7	**0.47**	**0.74**	**1.00**

8. WRS Algorithm

Weighted random sampling (WRS) is a class of algorithms for drawing items from a collection, where each item's probability of being selected is proportional to an assigned weight [2]. To construct the target training set $\mathcal{D}_{tgt}$ by resampling a source dataset $\mathcal{D}_{src}$, WRS effectively sample items with high-quality scores.

WRS first calculates the expected sampling count $r(v_i)$ for each data point $v_i \in \mathcal{D}_{src}$ using a softmax distribution over its quality score $Q(v_i)$, and scaled by the desired size of the target dataset $|\mathcal{D}_{tgt}|$:

$$r(v_i) = \frac{\exp(Q(v_i)/\tau)}{\sum_{j \in \mathcal{D}_{src}} \exp(Q(v_j)/\tau)} |\mathcal{D}_{tgt}| \tag{16}$$

where the temperature parameter $\tau > 0$, controls the sharpness of the sampling distribution.

The real-valued count $r(v_i)$ is then converted into an integer number of samples via stochastic rounding, ensuring the expected count for each item matches $r(v_i)$ [7].

9. Ablation Studies

9.1. Impact of Fusion Size M

The results in Table 8 demonstrate a clear correlation between the number of fused rPPG methods (M) and estimation accuracy. Relying on a single method ($M = 1$), results in the poorest performance. As we increase the fusion size from $M = 3$ to our chosen configuration of $M = 7$, we observe a consistent and significant reduction in both error metrics. This trend validates our core principle that a consensus-based fusion becomes more robust as it incorporates a larger set of diverse methods, effectively compensating for the idiosyncratic errors of individual rPPG estimators.

Scene-Level Noise Perception Template

Annotator Task: rPPG Video Data Quality Evaluation
Role: You are a Remote Photoplethysmography (rPPG) Training Data Annotator. Your job is to evaluate the quality of facial videos used for rPPG heart rate estimation.
Objective: Assess each video using the noise-based evaluation dimensions below. For each dimension, assign a score based on the provided criteria to determine the video's quality for rPPG analysis. Higher scores indicate lower noise impact (better quality).
Evaluation Dimensions:
1. Head Movement Noise (0–3 points)
- Evaluate: The presence, type, and severity of head/body motion and its effect on ROI tracking and potential pulse-frequency confounds. Distinguish benign actions (blinking, slight speech) from disruptive motion (rapid nodding/shaking/turning).
- Score:
- 0: Severe movement with frequent rapid rotations/translations, pronounced blur, or repeated ROI tracking failures.
- 1: Moderate movement or periodic motion likely to inject spurious frequencies or intermittently destabilize tracking.
- 2: Mild motion (brief expressions/speaking) with stable tracking and negligible blur.
- 3: Negligible motion; subject remains essentially still with no apparent periodic motion confounds.
2. Illumination Noise (0–3 points)
- Evaluate: Intensity, uniformity, and stability of facial lighting; presence of flicker, exposure clipping, shadows, or drift over time.
- Score:
- 0: Severe issues (strong flicker/strobing, under/overexposure with clipping, large brightness swings, heavy shadowing).
- 1: Moderate nonuniformity or temporal drift; some shadowing or localized saturation but face remains partly usable.
- 2: Mild variations; mostly uniform and stable lighting without clipping; minor fluctuations only.
- 3: Consistently uniform, flicker-free lighting; skin well exposed across the recording.
3. Skin-related Noise (0–2 points)
- Evaluate: Visibility of subtle skin chrominance changes given skin tone and texture; effects of specular highlights, facial hair, makeup, filters, or partial occlusions on usable skin ROI (forehead/cheeks).
- Score:
- 0: Poor visibility—very dark or saturated regions, strong glare, heavy makeup/filters/facial hair, or notable occlusions leaving minimal usable skin.
- 1: Moderate visibility—usable areas exist but exposure/texture is uneven or partially occluded.
- 2: Excellent visibility—proper exposure on cheeks/forehead with clear, subtle color variations and minimal occlusion.
4. Camera-related Noise (0–2 points)
- Evaluate: Sensor and encoding quality affecting rPPG (compression artifacts, color jitter, noise in low light, resolution/focus stability, frame-rate stability, rolling shutter).
- Score:
- 0: Heavy compression or pronounced artifacts/noise; low effective detail or unstable focus/frame rate that would significantly impair rPPG.
- 1: Moderate artifacts/noise; adequate detail with occasional autofocus/gain adjustments or minor instability.
- 2: Minimal artifacts; high-quality, stable capture with sufficient resolution and faithful color.
Requirements: Based on the above dimensions, score the video content, first stating the evaluation reasons for each dimension, then providing the quality assessment score. The final score is the sum of all dimensions, ranging from 0–10 points. Output format is JSON: { "Evaluation Reasons": { "Head Movement Noise": "...", "Illumination Noise": "...", "Skin-related Noise": "...", "Camera-related Noise": "..." }, "Scores": { "Head Movement Noise": X, "Illumination Noise": X, "Skin-related Noise": X, "Camera-related Noise": X, "Final Score": X } }
Note: Higher scores indicate better quality (lower noise impact) for rPPG applications. Videos scoring 6 or higher are generally suitable for accurate heart rate estimation.
Evaluate the following video:
<Video>

Figure 3. **Prompt for Qwen3-VL to assess scene-level quality.**

9.2. MLLM Generalization and Stability

We assessed generalization and stability by testing on Gemini 3.0 Pro and Kimi K2.5, using GPT-5.2 to generate perturbed prompt variations. Table 9 confirms our framework's robustness, showing minimal fluctuation in MAE, RMSE, and R across models and runs. Consequently, we retained Qwen3-VL as our primary model, as it provides the most favorable balance of latency and cost.

Table 9. **Generalization and stability analysis across different MLLMs.**

MLLM	MAE↓	RMSE↓	R↑
Qwen3-VL	0.46 ± 0.02	0.77 ± 0.06	1.00 ± 0.00
Gemini 3.0 Pro	0.52 ± 0.05	0.73 ± 0.08	1.00 ± 0.00
Kimi K2.5	0.55 ± 0.09	0.86 ± 0.11	1.00 ± 0.00

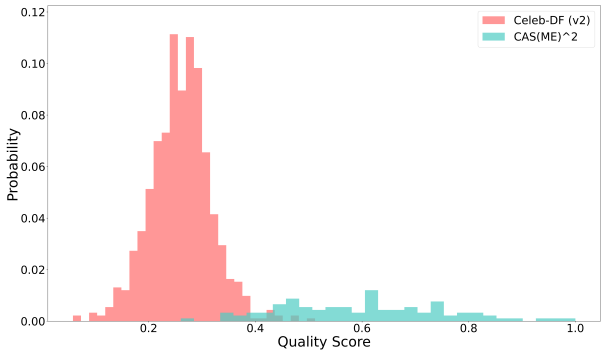

Figure 4. **Distribution of quality scores over the CAS(ME)^2 and Celeb-DF (v2) datasets.**

10. Visualization

Figure 4 illustrates a disparity in the quality score distributions between the CAS(ME)^2 [6] and Celeb-DF (v2) [5] datasets. Scores for CAS(ME)^2, a controlled dataset, are concentrated in the high-quality range (0.4-0.8), reflecting its consistent signal fidelity. In contrast, scores for the "in-the-wild" Celeb-DF (v2) are predominantly clustered in the lower 0.1-0.4 range. While the latter offers valuable scenic diversity, it is plagued by poor signal quality. This quantitative analysis validates our core premise: a fundamental trade-off exists between data quality and diversity, necessitating a robust curation strategy to effectively leverage unvetted datasets.

11. Failure Cases and Mitigation Mechanisms

11.1. Signal-level Branch Failure

Figure 5(a) illustrates a video from the "in-the-wild" MEVIEW dataset [4]. Existing estimators (GREEN, ICA, LGI, OMIT) produced inflated SNR values ranging from 16.85 to 26.69, leading to an erroneous consensus SNR of 20.72. This error likely stems from misinterpreting flickering background figures as physiological pulses. However, the scene-level branch effectively mitigated this by detecting the periodic visual noise, assigning a low quality score (6.00/10.00) to flag the sample.

(a) (b)

Figure 5. **Representative failure cases.** (a) Environmental noise interference; (b) AI-synthesized synthetic faces.

11.2. Scene-level Branch Failure

Figure 5(b) depicts an AI-synthesized video void of rPPG signals. Here, the scene-level branch incorrectly assigned a high quality score (9.00/10.00) due to the video's visual stability. The signal-level branch compensated for this misclassification, yielding a consensus SNR of -2.67 and correctly identifying the absence of physiological signals.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

[2] Pavlos S Efraimidis and Paul G Spirakis. Weighted random sampling with a reservoir. *Information processing letters*, 97 (5):181–185, 2006. 1

[3] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[4] Petr Husák, Jan Cech, and Jiří Matas. Spotting facial micro-expressions "in the wild". In *22nd Computer Vision Winter Workshop (Retz)*, pages 1–9, 2017. 3

[5] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020. 3

[6] Fangbing Qu, Su-Jing Wang, Wen-Jing Yan, He Li, Shuhang Wu, and Xiaolan Fu. Cas(me)^2: a database for spontaneous macro-expression and micro-expression spotting and recognition. *IEEE Transactions on Affective Computing*, 9(4):424–436, 2017. 3

[7] Xiangyu Xi, Deyang Kong, Jian Yang, Jiawei Yang, Zhengyu Chen, Wei Wang, Jingang Wang, Xunliang Cai, Shikun Zhang, and Wei Ye. Samplemix: A sample-wise pre-training data mixing strategey by coordinating data quality and diversity. *arXiv preprint arXiv:2503.01506*, 2025. 1