ICIP abu dhabi 2024
IEEE International Conference on Image Processing
27 - 30 October 2024, Abu Dhabi, UAE

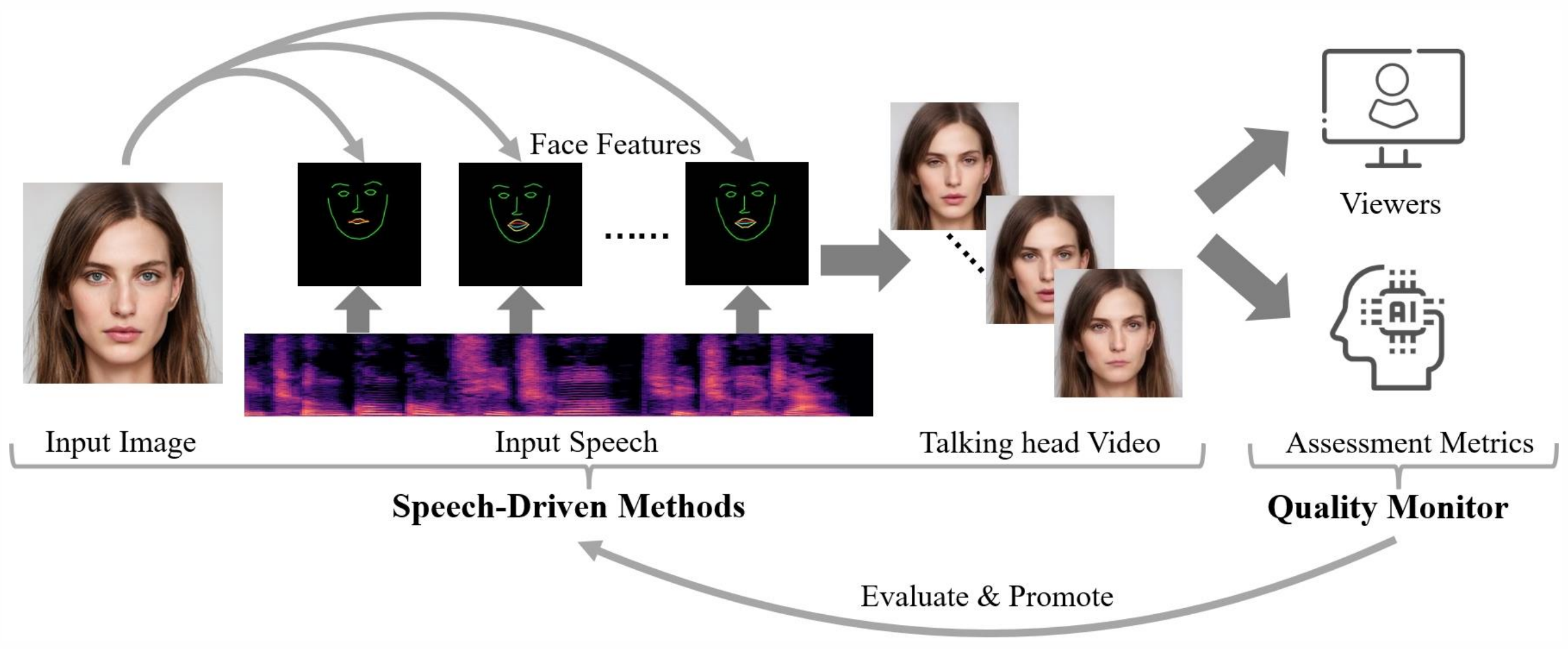
# THQA: A Perceptual Quality Assessment Database for Talking Heads

**Yingjie Zhou, Zicheng Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Zhihua Wang, Xiao-Ping Zhang, Guangtao Zhai**

**Shanghai Jiao Tong University，PengCheng Laboratory，Tsinghua University，Shenzhen MSU-BIT University**

Email: zyj2000@sjtu.edu.cn　　Project GitHub: https://github.com/zyj-2000/THQA

## Intro: Talking Heads & Quality



The process of designing digital human is a tedious and time-consuming, especially for parts like the human head with rich details and identity information. However, the advent of AI has simplified this process. The speech-driven methods are capable of generating the mouth shape of the character based on the input speech. However, despite the fact that multiple approaches have been proposed, this technology is currently immature and still suffers from many quality issues that seriously affect the user‘s visual experience. Therefore, **it is important to conduct quality assessment for this type of talking heads (THs).**

**Table 3**. Details of the speech-driving methods employed for the generation of TH videos.

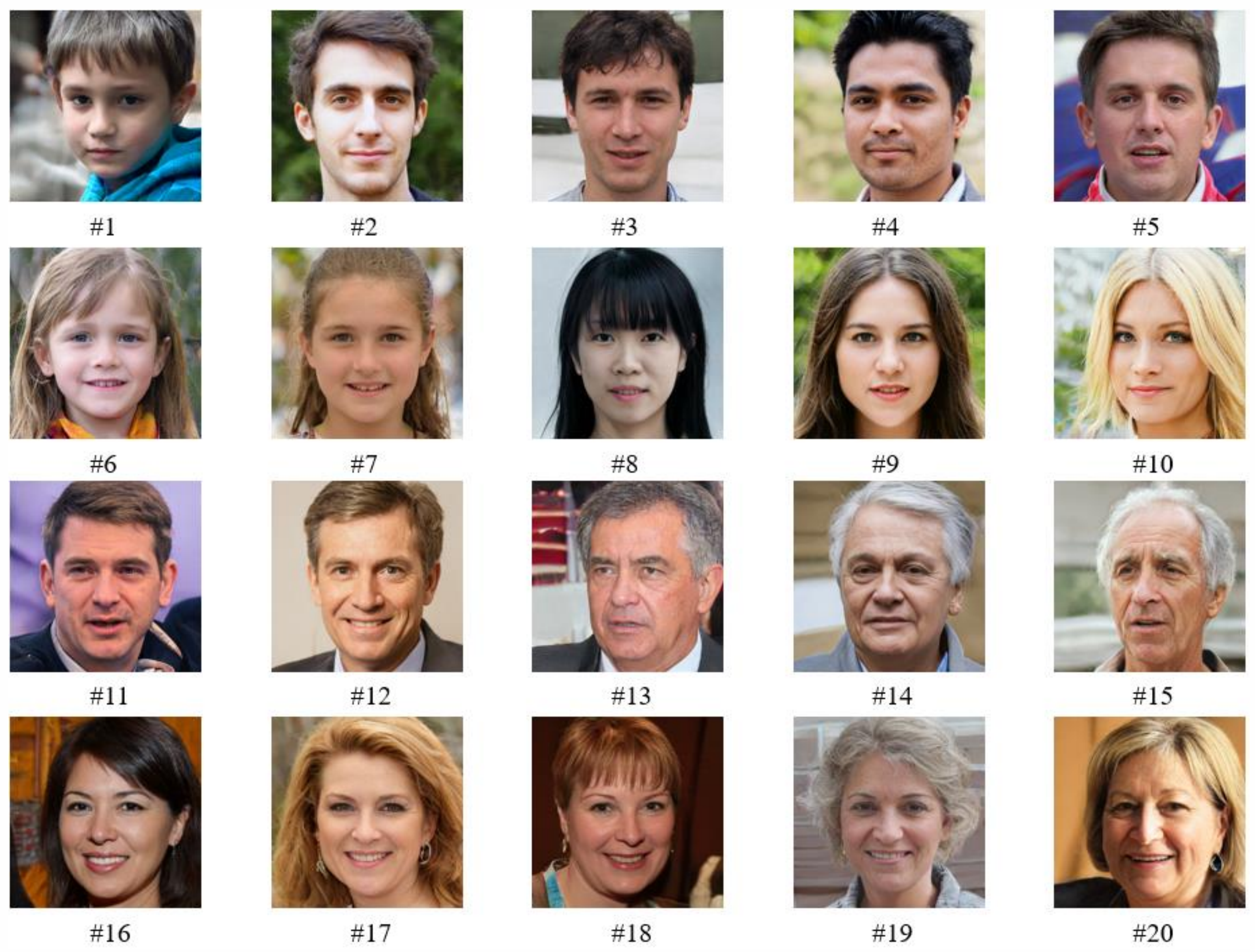
| Type | Label | Methods | Year | Head Motion | Output Resolution |
|---|---|---|---|---|---|
| Image-based | MT | MakeIttalk [12] | 2020 | ✓ | 256×256 |
| | AH | Audio2Head [13] | 2021 | ✓ | 256×256 |
| | ST | Sadtalker [11] | 2023 | ✓ | 512×512 |
| | DT | Dreamtalk [15] | 2023 | ✓ | 256×256 |
| Video-based | WL | Wav2Lip [16] | 2020 | ✕ | 1024×1024 |
| | VR | Video-Retalking [17] | 2022 | ✕ | 1024×1024 |
| | DN | DINet [18] | 2023 | ✕ | 1024×1024 |
| | IL | IP-LAP [14] | 2023 | ✕ | 1024×1024 |

## Insights & Benchmark

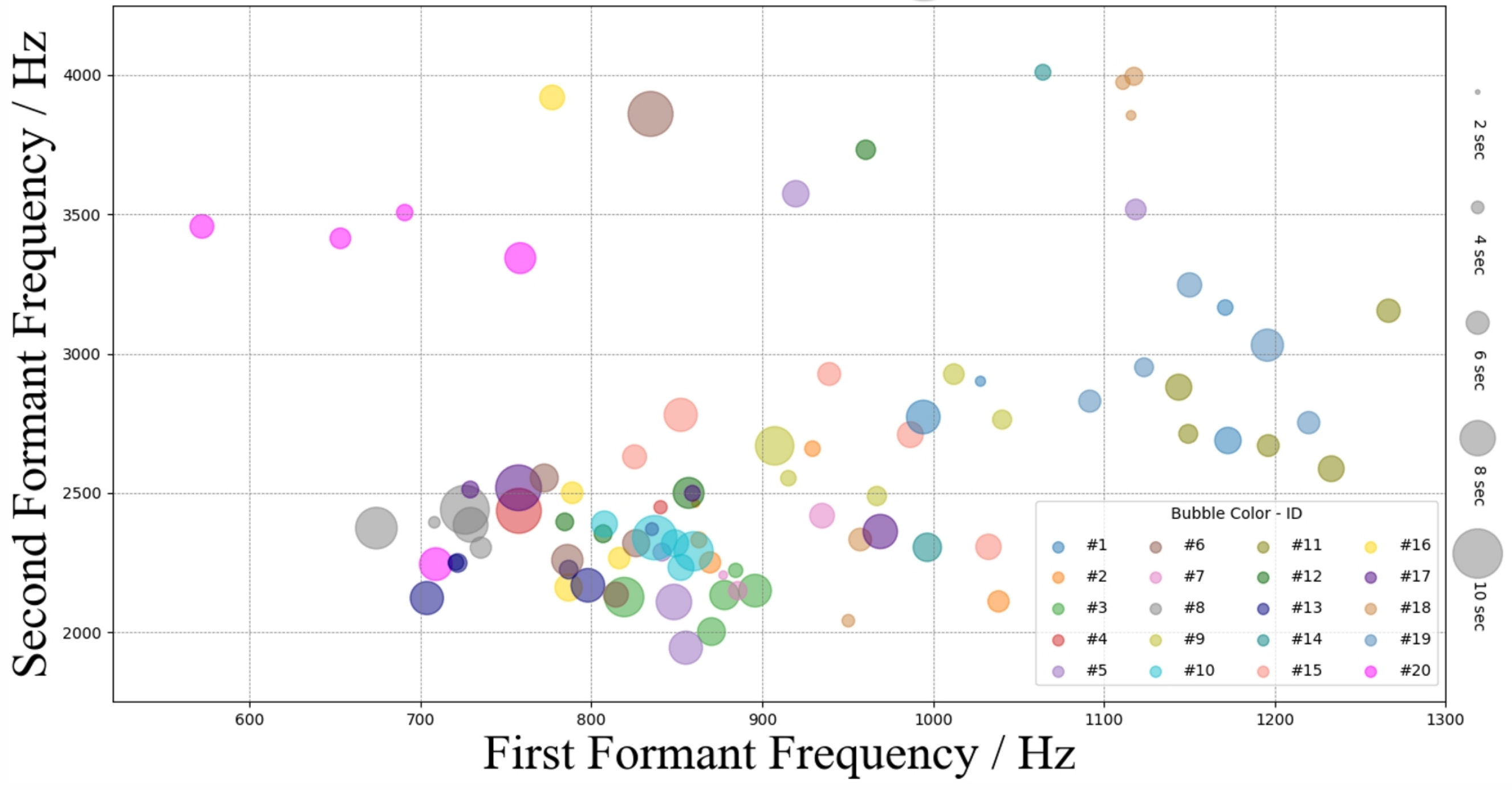
**Table 5**. Benchmark performance on the THQA database. Best in **bold.**

| Type | Method | SRCC↑ | PLCC↑ | KRCC↑ | RMSE↓ |
|---|---|---|---|---|---|
| IQA | BRISQUE | 0.4856 | 0.5970 | 0.3454 | 0.8227 |
| | NIQE | 0.1007 | 0.1389 | 0.0804 | 0.9673 |
| | IL-NIQE | 0.1199 | 0.1298 | 0.0979 | 0.9691 |
| | CPBD | 0.1184 | 0.1802 | 0.0932 | 0.9532 |
| VQA | VIIDEO | 0.1777 | 0.1891 | 0.1354 | 0.9595 |
| | V-BLIINDS | 0.4949 | 0.6403 | 0.3533 | 0.7976 |
| | TLVQM | 0.0254 | 0.0355 | 0.0209 | 1.0853 |
| | VIDEVAL | 0.0317 | 0.0358 | 0.0231 | 1.1916 |
| | VSFA | **0.7601** | **0.8106** | **0.5830** | **0.5966** |
| | RAPIQUE | 0.1789 | 0.1908 | 0.1277 | 1.0162 |
| | SimpVQA | 0.6800 | 0.7592 | 0.5052 | 0.6361 |
| | FAST-VQA | 0.6389 | 0.7441 | 0.4677 | 0.6983 |
| | BVQA | 0.7287 | 0.7985 | 0.5549 | 0.6094 |

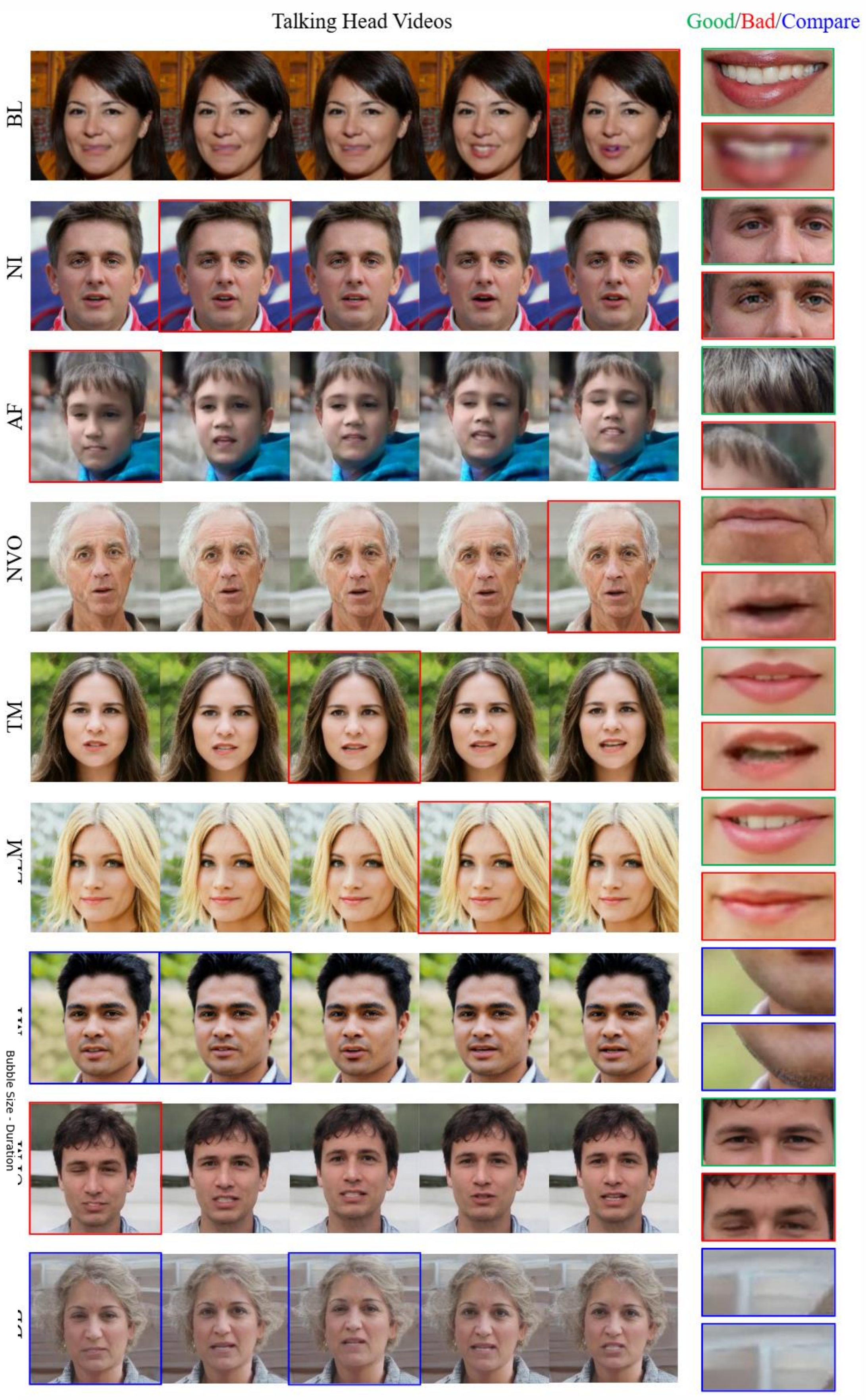
## THQA Database & Distortions



*Portrait Images*



*Acoustic Features*



*Distortions*

## Main parameters of the database and Access

- **20** portrait images with different gender and age
- **100** speeches with rich acoustic features
- **8** representative speech-driven talking head methods
- **800** talking head videos
- **9** common distortions

The experimental results are presented in Table, offering insights that lead to several noteworthy conclusions. 1) Notably, deep learning-based quality assessment methods, except RAPIQUE, exhibit superior performance compared to traditional methods reliant on manually extracted features. This superiority can be attributed to the inherent challenge of employing specific features for the quality assessment of AI-generated videos such as those in the TH domain. 2) VSFA emerges as the leading performer among all the compared methods for TH videos. However, it is important to note that despite achieving state-of-the-art performance, the evaluation results for TH videos still exhibit a discernible gap when compared to subjective human visual perception. 3) The performance evaluation of existing quality assessment methods on the THQA database underscores the limitations inherent in these methods for TH videos. Consequently, **it is evident that more effective and accurate assessment algorithms need to be explored and developed to address these limitations.**