ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888660

Subjective Fidelity Assessment of Audio- and Video-Driven Talking Head Generation Methods

Anthony Trioux*, Yusong Gao*, Jiarun Song, Wenjie Wu, Faming Ma, Fuzheng Yang
School of Telecommunications Engineering, Xidian University, Xi'an, China
{anthony_trioux, jrsong}@xidian.edu.cn ; {ysgao_1, wwj, mfm}@stu.xidian.edu.cn ; fzhyang@mail.xidian.edu.cn

***Abstract*—Audio- and Video-Driven Talking Head Generation methods have attracted considerable research interest due to recent advances in Artificial Intelligence Generated Content (AIGC) technologies. In such approaches, a single image is artificially animated by leveraging audio and/or motion features extracted from video sources. Despite notable progress, current performance assessments rely primarily on traditional objective metrics, often neglecting subjective evaluation aspects. To address this issue, we propose in this paper a subjective *fidelity* assessment of recent Audio- and/or Video-Driven Talking Head Generation methods. This study aims to assess how accurately and convincingly the generated video reproduces the visual and behavioral characteristics of a real human face, as well as how closely the video aligns with expected natural human expressions, movements, and/or audio synchronization. In order to provide a detailed assessment of the *fidelity* in the context of talking heads, our study focuses on six key criteria: Overall Fidelity, Gaze Fidelity, Audio-Video Sync Fidelity, Head Pose Fidelity, Expression Fidelity, and Overall Visual Quality. Experiments results reveal a nuanced picture of the fidelity in this context, where the performance varies significantly depending on the video content itself as well as how the animation is generated, highlighting the needs for further research. This research represents an initial step towards the evaluation of Audio- and Video-Driven generative image animation methods for Talking heads while offering insights for improving the accuracy and realism of those techniques. The dataset and corresponding results are available at https://github.com/a-trioux/Subjective-Fidelity-Assessment-Talking-Head.**

***Index Terms*—Audio- and Video-Driven Talking Head Generation, Multimodal Fidelity Assessment, AIGC, Generative models, Video conferencing**

I. Introduction

In recent years, deep generative models have seen rapid advances and widespread market adoption, particularly within the field of Artificial Intelligence Generated Content (AIGC) [1], [2]. These models enable the creation of realistic visual and audio content, with a notable application in talking head generation, where single facial images are animated based on audio inputs (Audio-Driven) and/or video inputs (Video-Driven). Leveraging state-of-the-art neural architectures, such as Generative Adversarial Networks (GANs) [3] and Variational AutoEncoders (VAEs) [4], AIGC has enabled more immersive and engaging virtual experiences, including applications like (virtual) video conferencing, telepresence, interactive media, etc.

Generative Face Video Coding (GFVC) has emerged as a key research area within this field. By integrating generative models into video coding frameworks, GFVC efficiently encodes and reconstructs facial video content, significantly reducing bandwidth requirements while enhancing visual quality [5]–[7]. This is particularly valuable in bandwidth-constrained scenarios, such as congested networks or areas with weak radio coverage, where smooth transmission and high-quality visual representation are essential for user experience. Recognizing its potential, the Joint Video Exploration Team (JVET) recently established an Ad hoc Group on GFVC to investigate its viability for real-world applications, including software implementation, test conditions, coordinated experimentation, interoperability with existing codecs, etc. [7].

Although Talking Head Generation methods have shown impressive progress, challenges remain in achieving consistently high fidelity across diverse content types, particularly for attributes such as gaze direction, head pose, and lip synchronization. In addition, current evaluations predominantly rely on objective metrics such as Peak Signal-to-Noise Ratio (PSNR) [8], SSIM (Structural Similarity Index) [9], etc. which often fail to align with human perception. Recently proposed metrics, such as LPIPS [10] and DISTS [11], offer better alignment with human visual perception but are not specifically tailored to the unique requirements of talking heads or video conferencing applications [12]. To address these limitations, psychovisual studies [12], [13] have started to assess the subjective quality of generated talking heads, emphasizing the importance of assessing the perceived realism and quality of generated talking heads.

Our work aims to build upon these previous studies, while being distinct, by providing a detailed subjective assessment framework that considers multiple dimensions of the *Fidelity*, including Overall Fidelity, Expression Fidelity, Audio-Video Synchronization, Head Pose Fidelity, Gaze Fidelity, and Overall Visual Quality. ***Fidelity*** in this context refers to the degree to which the generated talking head video accurately and convincingly reproduces the visual and behavioral characteristics of a real human face. This includes how well the generated video replicates expected natural human expressions, movements, and video-audio synchronization.

The remainder of the paper is organized as follow. Section II reviews recent audio- and video-driven based approaches used in this study and introduces key parameters that influence the received quality in a talking head context. The experimental setup of the subjective evaluation is explained in Section III. Results are shown in Section IV. Conclusions and discussions are presented in Section V.

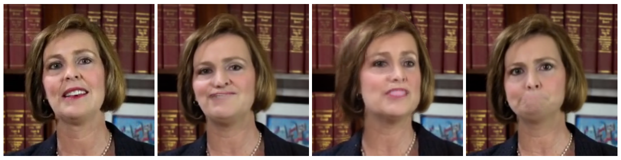
*equal contribution, corresponding author: anthony_trioux@xidian.edu.cn

II. Selected Methods and Evaluation Criteria

In this study, we select the three following methods for our subjective assessment of talking head generation: EDTalk [14], FOMM [15], and LivePortrait [16]. These methods are chosen based on their current relevance and performance in the field. EDTalk and LivePortrait represent recent advances in the literature, demonstrating strong performance in talking head synthesis, while FOMM is notable for its use in ongoing JVET standardization efforts. **EDTalk** [14] introduces an efficient disentanglement framework that leverages an Audio-Motion Module to extract pose, mouth, and expression parameters from audio. This framework also allows to enhance the accuracy of talking head synthesis by enabling the extraction of pose parameters from video and synchronizing the generated face with the head posture of the original video. Despite its advancements, EDTalk still faces challenges in generating realistic head pose and gaze direction as shown in Fig. 1. **FOMM** [15] is an image-driven method that learns implicit keypoints and their corresponding Jacobian matrices through unsupervised learning, so that local affine transformations can be used to represent the motion in the neighborhood of each keypoint. Finally, motion optical flow is used to warp the source/reference image to generate a video sequence. Finally, **Liveportrait** [16] relying on FV2V [17], incorporates advanced parameters such as 3D keypoints and pose parameters to better manage head movements. By using landmarks from the eyes and lips as guidance, LivePortrait improves the generation of detailed and complex expressions. However, all the above methods still face challenges to generate realistic lips and/or precise gaze direction, as illustrated in Fig. 1.

To comprehensively evaluate these methods, we propose a subjective assessment framework based on six key criteria: Overall Fidelity, Gaze Fidelity, Audio-Video Sync Fidelity, Head Pose Fidelity, Expression Fidelity, and Overall Visual Quality. This selection is based on the limitations observed in the current methods:

- Overall Fidelity addresses the overall accuracy and convincing naturalness of the generated video in reproducing real human characteristics.
- Gaze Fidelity evaluates the precision of gaze direction, which is crucial for natural interactions and often problematic in existing methods.
- Audio-Video Sync Fidelity measures the alignment between audio and visual components, which remains a challenge in dynamic scenarios.
- Head Pose Fidelity assesses how well the generated head movements match the intended posture, a known issue for video content with rapid pose changes.
- Expression Fidelity focuses on the naturalness of facial expressions, a critical aspect that can be affected by occlusions and disocclusions.
- Overall Visual Quality encompasses the general visual appeal of the generated video, including the presence of artifacts (blurriness, unnatural textures, etc.) and their effect on the perceived quality.



(a) Orig. Frame 58 (b) FOMM [15] (c) EDTalk [14] (d) LivePort. [16]

Fig. 1. Illustration of artifacts generated by the considered models on video sequences from the HDTF database [18]. (a) Original image, (b) FOMM [15], (c) EDTalk [14], (d) LivePortrait [16].

These criteria are selected to address the specific challenges highlighted in the limitations of Audio- and Video-driven generative methods, providing a comprehensive evaluation framework that targets the key aspects of *fidelity* in talking head contexts.

III. Subjective fidelity evaluation

Environment: The test was carried out according to the ITU-R BT.500-14 recommendation [19], in a dark and quiet room with controlled ambient luminance and color temperature. The screen used for display is a 27”, 2K resolution Dell SE2723DS.

Observers: A total of 27 participants, including 18 men and 9 women, took part in the study. Their ages ranged from 22 to 26, with a mean age of 22.74. All participants had either normal or corrected to normal visual acuity. The panel was composed of both experts (less than 15%, familiar with AIGC methods or having specific knowledge in video coding) and newcomers to the field.

Methodology: Our test followed an Absolute Category Rating (ACR)-like method [20], where participants were asked to evaluate the six key criteria on a 5-point Likert scale. Prior to the actual test, participants were familiarized with the testing environment through a pre-training session, with two video content not included in the test itself. Participants were then let alone and randomly presented with video stimuli, featuring various generative methods including: EDTalk [14], EDTalk+pose [14], FOMM [15], and LivePortrait [16]. Additionally, the original video was included in the test set to serve as a baseline, allowing an assessment of how the original content performed across the six criteria. This also acted as a defense mechanism against outliers, as it is highly unlikely that these high-quality videos would receive low scores. Note that a VVC-compressed [21] version of the videos was also included in the test to get an idea on how traditional codecs perform in this context of ultra-low bitrate video conferencing/talking heads. During the test, participants were allowed to replay the video stimuli if necessary, ensuring a thorough evaluation. Each participant evaluated a total of 84 stimuli, with a randomized list generated for each person, ensuring no consecutive repetition of the same content. The entire test took approximately one hour and participants were given the option to rest after evaluating each stimulus. In addition, a 20-minute break was included during the test to avoid fatigue, according to the ITU recommendations [22].

Material selection: For this study, a hundred of video-conferencing-like videos containing audio were carefully selected on YouTube based on copyright reuse. YouTube was chosen as a default-source due to the limited availability of publicly accessible talking-head video datasets including audio. Each video was cropped to focus on the talking-head content, with an average duration of each stimulus of about 18 seconds, giving participants enough duration time to assess each of the six criteria. From the initial pool, 16 videos were selected after a thorough pre-screening process conducted by two experts in the field. These videos were chosen to represent a wide range of fidelity assessment challenges, carefully selected to focus on high-quality content. Those experts did not participate in the subjective tests. A particular attention was made to audio and video diversity, ensuring representation of different ethnicities, low-to high speed speeches, a balanced male-to-female ratio, and individuals with accessories such as headphones and glasses. Additionally, the spatio-temporal complexity of each video was evaluated using the Spatial Information (SI) and Temporal Information (TI) indexes [22] as well as the newest Video Complexity Assessment (VCA) [23] metric. The scatter plots of SI, TI, and VCA values for the selected stimuli are available in the GitHub link. Note that we use those indicators with caution as talking-head are different from traditional content and might necessitate the needs of a specific complexity metric dedicated to such content. This is out of the scope of the paper and is left as future work.

IV. Experimental results

A. Bitrate consideration

Except FOMM [15] which is currently under investigation in JVET standardization group, other methods were not originally designed to handle bitrate constraints. However, in practical videoconferencing applications, compression is required for efficient transmission. To address this, we adopted the encoding mechanism utilized by the GFVC Adhoc group of JVET to compress animation features such as keypoints and pose [24]. Specifically, a residual computation between current and previous frames is firstly performed. Then a scalar quantization is used as follows,

$$\tilde{x} = \lfloor x \times QS \rceil, \tag{1}$$

where QS denotes the Quality Scale, x denotes the original data, $\lfloor \cdot \rceil$ denotes the rounding operation to the nearest integer and $\tilde{x}$ denotes the data after quantization. Finally, a variable length encoding based on Exp-Golomb codes is used to obtain the compressed data in binary format.

Audio was encoded using the libavformat [25], while the first (reference) frame was compressed with the JPEG codec using a high quality factor ($QF = 90$) to ensure very low distortion on the reference image. The resulting average bitrate to encode the video part of each method is as follow: FOMM ($QS = 256$) = 9.55kbps, EDTalk+pose ($QS = 1024$) =5.85kbps, Liveportrait ($QS = 1024$) = 8.44kbps. Note that FOMM data is more severly quantized compared to the other methods as it requires to transmit for each encoded keypoint, an additional 2×2 jacobian matrix to perform the animation [15]. In addition, since EDTalk is an audio-driven approach, it only requires the transmission of a single frame at the beginning. Therefore the notion of encoding video bitrate does not really apply in that case. For VVC, we selected an average value of 8.82kbps, in order to offer *viable* results while being as close as possible to the above bitrate.

B. Overall results

Prior to data analysis, the ITU-R BT.500-14 outlier detection method [19] was applied to filter out observers with inconsistent voting patterns. No outliers were detected. Next, we calculated the average scores for each of the six assessment criteria across all stimuli and methods (four generative models, the original videos, and the VVC-encoded ones). Given the limited number of participants, Confidence Intervals (CI) were also computed as recommended in [19]. The average scores shown in the radar chart of Fig. 2 reveal relatively narrow CIs, indicating consistent results across participants. As anticipated, original achieved the highest scores, while VVC produced the lowest scores due to limited/low bitrate. Conversely, the distinction between the remaining methods is more complex. While these newer techniques generally obtained low-to-moderate scores across all criteria, some key differences emerged. For example, in gaze fidelity, EDTalk, an audio-driven only method, underperformed relative to the other generative methods due to its lack of motion features.

Each evaluation criterion is analyzed independently to ensure a comprehensive understanding of the strengths and weaknesses of the methods across different fidelity dimensions. While the ”Overall Fidelity” and ”Overall Visual Quality” metrics may intuitively be correlated with the other specific fidelity dimensions (e.g., Gaze Fidelity or Expression Fidelity), their exact relationship requires further investigation. Future work could involve a detailed correlation analysis to explore these interdependencies, as well as potential methods to integrate these criteria into a unified assessment metric using weighting schemes or statistical models.

The results also highlight further areas in the talking-head domain that require additional research, particularly in addressing rapid pose changes, accurate gaze alignment, audio-video synchronization, and natural expression fidelity.

C. Per-sequence results

Assessing the evaluation criteria based on an average over fourteen sequences may not fully capture the nuances of the results. To address this, we present radar charts in Fig. 3 and visual illustrations in Fig. 4 for two selected sequences with distinct characteristics. Video 1 offers in average higher scores due to low head pose changes and uniform background. Video 2 represents a high-complexity content due to rapid eye, mouth and pose changes as well as complex background. This results in lower scores in general, with particular low values for Liveportrait, especially regarding the Audio-Video sync due to poor mouth reconstruction as shown in Fig. 4.

The results presented in this paper highlight several limitations in current audio- and video-driven generative approaches that should be addressed in future work. For instance, face-background entanglement remains a key challenge that is

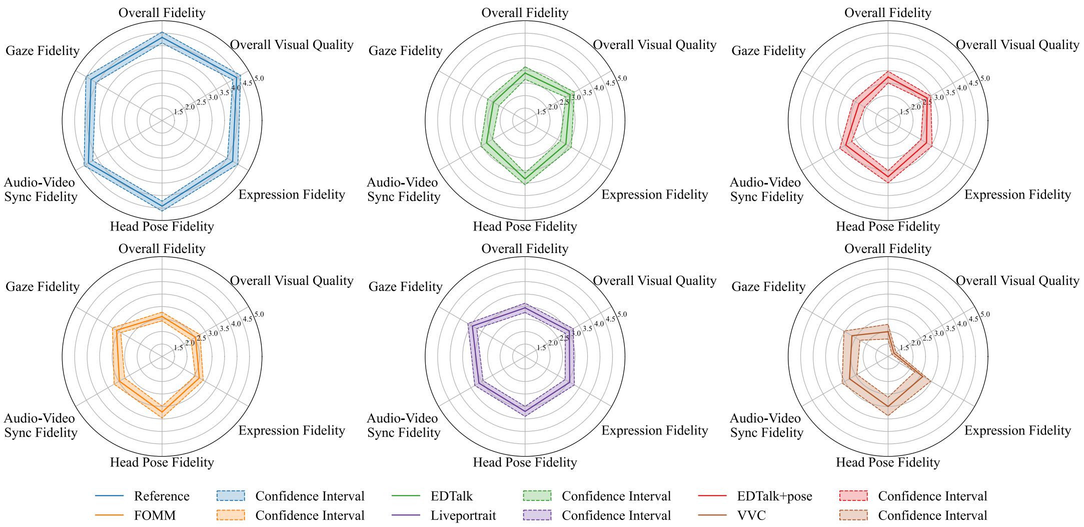


Fig. 2. Illustration of the radar chart of the subjective assessment. From top to bottom and left to right: Reference (blue), EDTalk(green), EDTalk+pose (red), FOMM (orange), Liveportrait (purple) and VVC (brown). Solid line: Average Score, Dotted line: Confidence Interval.

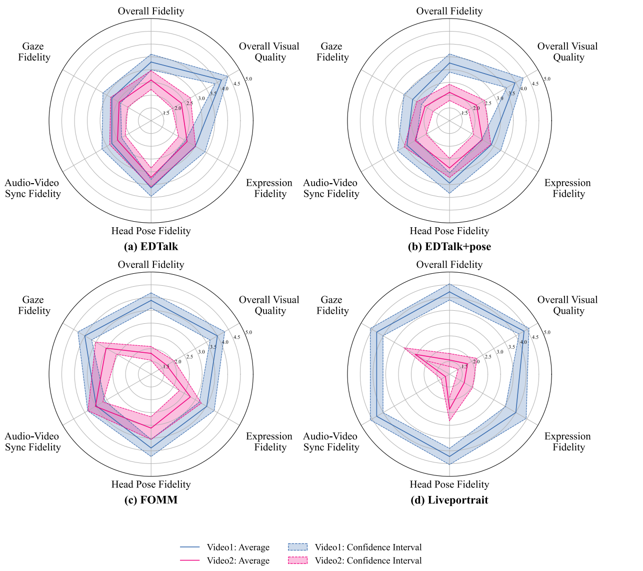
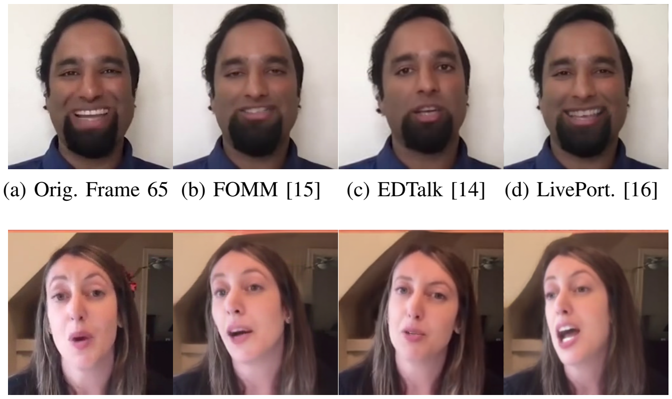


Fig. 3. Illustration of the radar chart on Video1 (blue) and Video2 (pink) from our dataset. From top to bottom and left to right: (a) EDTalk, (b) EDTalk+pose, (c) FOMM and (d) Liveportrait.

under investigation by the JVET standardization group [26]. Additionally, the discrepancy in results between videos also suggests potential weaknesses in the training process, specifically regarding how training video sequences should be selected to account for diverse scenarios.

V. Discussion and Conclusion

In this paper, we study the subjective fidelity assessment of recent audio- and video-driven generative methods in the context of talking-head videos. Our results provide a detailed view of the *fidelity* within this context, showcasing variation in performance based on both the video content and the animation generation methods (audio or video-based). These results underscore the necessity for further investigation. This study provides a preliminary evaluation of generative image animation techniques, offering insights to improve the accuracy and realism of these methods. Despite advances, current techniques still struggle with issues such as gaze direction and audio-video synchronization, etc. Future research may concern the assessment of popular objective metrics and their ability to predict the fidelity scores in this specific context as well as a deepen study on the impact of compression of the different elements (reference frame, features, audio) on the fidelity.



(a) Orig. Frame 65 (b) FOMM [15] (c) EDTalk [14] (d) LivePort. [16]

(e) Orig. Frame 413 (f) FOMM [15] (g) EDTalk [14] (h) LivePort. [16]

Fig. 4. Illustration of observed artifacts (entanglement, etc.) on the dataset used in this study. Top: Video1, Bottom: Video2. (a), (e) Original images, (b), (f) FOMM [15], (c), (g) EDTalk [14], (d), (h) LivePortrait [16].

Acknowledgments

This work was partly supported by the Department of Science and Technology of Shaanxi Province, China (Grant QCYRCXM-2023-044), by the State Administration of Foreign Experts Affairs, China (Grant Y20240269) and by the State Grid Corporation Headquarters Technology, China (Grant 5500-202424166A-1-1-ZN).

References

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, “A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT,” *arXiv preprint arXiv:2303.04226*, 2023.

[2] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li, M. Zhang, S. K. Dam, C. M. Thwal, Y. L. Tun, L. L. Huy *et al.*, “A complete survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 all you need?” *arXiv preprint arXiv:2303.11717*, 2023.

[3] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[4] L. Mescheder, S. Nowozin, and A. Geiger, “Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks,” in *International conference on machine learning*. PMLR, 2017, pp. 2391–2400.

[5] B. Chen, J. Chen, S. Wang, and Y. Ye, “Generative face video coding techniques and standardization efforts: A review,” in *2024 Data Compression Conference (DCC)*. IEEE, 2024, pp. 103–112.

[6] Z. Chen, H. Sun, L. Zhang, and F. Zhang, “Survey on visual signal coding and processing with generative models: Technologies, standards and optimization,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2024.

[7] B. Chen, Y. Ye, G. Konuko, G. Valenzise, S. Yin, and S. Wang, “AHG 16: Updated common software tools for generative face video compression,” in *The Joint Video Experts Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29, doc. no. JVET-AH0114*, 2024.

[8] Q. Huynh-Thu and M. Ghanbari, “Scope of validity of psnr in image/video quality assessment,” *Electronics letters*, vol. 44, no. 13, pp. 800–801, 2008.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018, pp. 586–595.

[11] J. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.

[12] B. C. M. W. Yixuan Li, Bolin Chen and S. Wang, “Perceptual quality assessment of face video compression: A benchmark and an effective method,” *arXiv preprint arXiv:2304.07056*, 2023.

[13] W. Zhang, C. Zhu, J. Gao, Y. Yan, G. Zhai, and X. Yang, “A comparative study of perceptual quality metrics for audio-driven talking head videos,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024.

[14] S. Tan, B. Ji, M. Bi, and Y. Pan, “EDTalk: Efficient disentanglement for emotional talking head synthesis,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[15] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” *Advances in neural information processing systems (NeurIPS)*, vol. 32, 2019.

[16] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, “LivePortrait: Efficient portrait animation with stitching and retargeting control,” *arXiv preprint arXiv:2407.03168*, 2024.

[17] T.-C. Wang, A. Mallya, and M.-Y. Liu, “One-shot free-view neural talking-head synthesis for video conferencing,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 10 039–10 049.

[18] Z. Zhang, L. Li, Y. Ding, and C. Fan, “Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[19] International Telecommunication Union, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Tech. Rep. ITU-R BT.500-14, 2019. [Online]. Available: https://www.itu.int/rec/R-REC-BT.500-14-201210-I/en

[20] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, “Performance comparisons of subjective quality assessment methods for mobile video,” in *2010 Second international workshop on quality of multimedia experience (QoMEX)*. IEEE, 2010, pp. 82–87.

[21] B. Bross, X. Wang, J. Ma, M. Chen, J.-R. Ohm, G. J. Sullivan, X. Liao, S. Liu, Z. Li *et al.*, “Versatile video coding (VVC) - specification,” ITU-T and ISO/IEC JTC 1, Tech. Rep. ISO/IEC 23090-3:2020 / ITU-T H.266, 2020. [Online]. Available: https://www.itu.int/rec/T-REC-H.266

[22] International Telecommunication Union, “Recommendation p.910: Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, Tech. Rep. T-REC-P.910-202310-I!!PDF-E, 2023. [Online]. Available: https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.910-202310-I!!PDF-E&type=items

[23] V. V. Menon, C. Feldmann, H. Amirpour, M. Ghanbari, and C. Timmerer, “VCA: video complexity analyzer,” in *Proceedings of the 13th ACM multimedia systems conference*, 2022, pp. 259–264.

[24] S. W. Golomb, “Run-length encodings,” *IEEE Transactions on Information Theory*, vol. 12, no. 3, pp. 399–401, 1966.

[25] L. Developers, *Libavformat Documentation*, Libav, 2024. [Online]. Available: https://libav.org/documentation/

[26] S. Gehlot, G.-M. Su, P. Yin, S. McCarthy, and G. J. Sullivan, “AHG9/AHG16: Showcase for picture fusion for generative face video sei message,” Joint Video Exploration Team (JVET), Rennes, France, Tech. Rep. JVET-AH0118, April 2024.