THEval. Evaluation Framework for Talking Head Video Generation

Supplementary Material

A. User study

To ensure a fair and unbiased comparison between methods, the algorithm selects videos randomly from common videos of all methods. By choosing videos at random, the evaluation avoids over-representing any specific content or scenario, which could otherwise skew results. Additionally, the algorithm randomly assigns each method to the left or right position in the user interface for each comparison to mitigates positional bias, ensuring that participants preferences are based on the video quality itself rather than the side on which it appears. Screenshot of the website is available on Figure 5. Each human rater has seen a total of 153 videos for all pairs of videos comprising the TH methods but also the real videos. We use Krippendorff's α to measures agreement across annotators. Using this metric on all 153 method-pair comparisons, we obtain $\alpha = 0.74$, indicating substantial agreement among participants.

B. Additional experiments on Syncnet instability

Our experiments showed that Syncnet LSE-C and LSE-D can be influenced by the way audio and video are encoded. Indeed, when changing the audio encoding from **mp4a** to **mpga**, the LSE-D and LSE-C vary significantly. When tested on the entire HDTF dataset, we notice that the average absolute difference in LSE-D and LSE-C between videos with **mp4a** or **mpga** audio is 0.4. This absolute difference can even reach values as high as 1.2 for some samples. We observe similar results when comparing video using **H.264** and **H.265** encodings. In both experiments there are no noticeable qualitative differences from a human evaluation standpoint. This confirms the findings of [49] that Syncnet is not stable and can be influenced by various factor unrelated to lips synchronization.

C. Temporal Drift in OmniAvatar

Figure 6. Temporal drift in OmniAvatar outputs over time. Ten frames are shown, sampled every 75 frames, illustrating gradual facial identity divergence, increasing visual artifacts, and the emergence of a color cast in longer videos.

D. State-of-the-art TH-generation methods used for benchmarking

Controltalk [57] is a talking face generation method to control face expression deformation based on driven audio, constructing the head pose and facial expression (lip motion) for both single image or sequential video inputs in a unified manner.

DaGAN [16] is a Depth-aware Generative Adversarial Network that first recovers dense 3D facial geometry (i.e., depth maps) from face videos in a self-supervised manner. This depth information is then used to guide the estimation of sparse facial keypoints and to learn a 3D-aware cross-modal attention mechanism, improving the generation of accurate face structures and motion fields.

Dimitra [7] is a diffusion based framework for TH generation that uses 3D motions as an intermediate step. It leverages audio features, phonemes and text to generate fully animated, realistic TH videos.

EchoMimic [5] uses audio speech to drive landmark sequences and employs a Latent Diffusion Model to convert input images into an efficient latent representation that is driven with the landmark sequence. It generates realistic results at high resolution.

EMOPortraits [10] builds upon the MegaPortraits model to enhance its capability for rendering intense and asymmetric facial expressions. It introduces architectural changes and a new training pipeline, including a novel dataset with extreme emotions (FEED), and incorporates a speech-driven mode, making it a multimodal framework for high-fidelity avatar animation.

Hallo2 [9] generates long-duration, high-resolution audio-driven portrait animations. It uses a patch-drop technique for temporal consistency, vector quantization for high resolution, and supports textual prompts for expression control.

LIA [39] is a self-supervised autoencoder that animates images by linear navigation in its latent space, removing the need for explicit structure representation. Motion is constructed by the linear displacement of latent codes, using a learned set of orthogonal motion directions.

LIA-X [42] is an interpretable portrait animator designed as an autoencoder that models motion transfer as a linear navigation of motion codes. It incorporates a Sparse Motion Dictionary to disentangle facial dynamics into interpretable factors, enabling a controllable 'edit-warp-render' strategy for precise manipulation of facial semantics.

Liveportrait [12] is an efficient video-driven portrait animation framework using an implicit-keypoint-based approach for good generalization and controllability. It features stitching and retargeting modules for precise control over elements including eye and lip movements with minimal computational cost.

MCNet [15] proposes a Memory Compensation Network to address ambiguities from dramatic motions in talking head generation. It learns a global facial meta-memory bank that provides structure and appearance priors. An implicit identity representation, learned from keypoints and features of the source image, is used to query this memory bank and compensate for warped features, particularly in occluded regions.

FOM [31] proposes novel motion representations for animating articulated objects by identifying and tracking object parts as regions rather than keypoints. In a fully unsupervised manner, it infers motion from the principal axes of these regions, disentangles shape and pose to prevent identity leakage, and models global background motion separately with an affine transformation.

OmniAvatar [11] is an audio-driven video generation model focused on creating full-body animations with adaptive and natural movements. It employs a LoRA-based training approach on a foundation model and introduces a multi-hierarchical, pixel-wise audio embedding strategy to enhance lip-sync accuracy and ensure audio features guide the entire body motion, not just the face.

Real3DPortrait [51] is a framework for realistic 3D talking portrait synthesis. It improves 3D reconstruction by distilling knowledge from a 3D face generative model into an image-to-plane network. It facilitates animation with a motion adapter and synthesizes a complete, realistic video by individually modeling the head, torso, and background, supporting both audio and video-driven inputs.

SadTalker [54] is a method for generating TH that produces realistic 3D motion coefficients for animated, audio-driven TH from a single image. It leverages full-image animation capabilities and utilizes pre-trained models to enhance the expressiveness and authenticity of the animated TH.

Wav2Lip [27] is a lip synchronization model for videos, aligning lip movements with audio segments for different identities in various settings. It uses a lip-sync discriminator based on Syncnet to enhance the precision of lip movements in TH videos. Wav2Lips does not generate the entire TH but only the mouth region. The generated mouth region is then integrated into the original video without altering the rest of the content.

X-Portrait [44] is a conditional diffusion model for expressive portrait animation. It uses a pre-trained Stable Diffusion model as a rendering backbone and achieves fine-grained motion control via ControlNet. It interprets dynamics directly from the raw driving video (implicit control) rather than relying on intermediate representations such as landmarks, and uses a cross-identity training scheme to mitigate identity leakage from the driver.

E. THEval Dataset

We curate a dataset comprising of TH-videos that includes multiple languages, varied clip lengths, which we have collected from 31 video channels comprising diverse speakers and settings such as lighting conditions, background, and camera angles. Figure 7 illustrates details on the dataset composition.

The dataset contains a total of 5,011 clips across six different languages to test the model's robustness and generalization capability beyond English-centric content. In particular, English is the most represented language with 1,680 clips, French is included in 890 clips, Chinese in 733, Spanish in 719, Japanese in 645, and Italian in 344 clips.

The dataset features a wide range of clip durations, from 2 seconds to 334.6 seconds, with the majority of clips being short. The first bin in (Center) Fig 7 corresponds to (0-10 seconds), accounting for a large percentage of the total clips, which is a typical for many TH datasets.

F. Existing metrics for evaluation of TH-generation

In this section we elaborate on the existing metrics for video evaluation, as well as on their limitations.

FID: The Fréchet Inception Distance (FID) constitutes an improvement of the Inception Score (IS). is a metric designed to evaluate the quality of generated images or videos. FID is computed by first extracting features from real and generated images using an inception network. Then, the features are treated as samples from two multivariate Gaussian

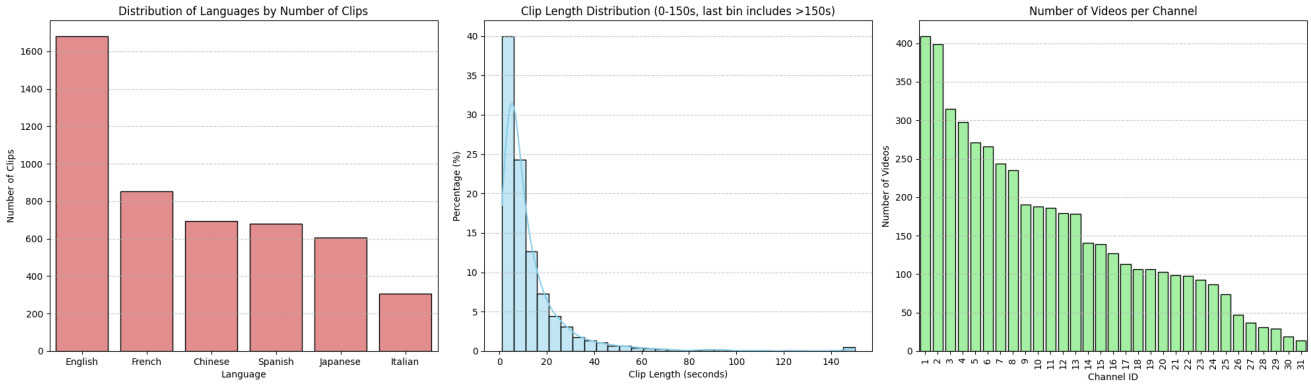

Figure 7. **Dataset Distribution Details.** (Left) Distribution of the 5,011 clips by language. (Center) Distribution of clip lengths, showcasing a high concentration of short clips. (Right) Distribution of videos across the 31 unique source channels, illustrating channel diversity.

distributions (real and generated) and the Fréchet distance between the two distributions is computed.

The Fréchet distance measures the distance between a generated image set and a source dataset, and is calculated as

$$\text{FID} = ||\mu_r - \mu_g||_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \quad (13)$$

where μ_r and μ_g are the mean feature vectors of the real and generated images respectively, Σ_r and Σ_g are the covariance matrices of the real and generated images respectively.

FID is highly dependent on the performance of inception network and assumes that the images features follow a Gaussian distribution which might not be true. FID is also biased when evaluated on a finite set due to the assumption of Gaussian distribution [6]. To be accurate, FID must be evaluated on a set that is large enough which might not be possible for all generation tasks. When used for video evaluation FID will only evaluate independent frame quality without regards for the temporal coherency.

FVD: The Fréchet Video Distance (FVD) is similar to FID but uses a network adapted for videos to extract the features. FVD is calculated as

$$\text{FVD} = ||\mu_r - \mu_g||_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}, \quad (14)$$

where μ_r and μ_g are the mean feature vectors of the real and generated videos respectively, Σ_r and Σ_g are the covariance matrices of the real and generated videos respectively.

FVD has the same limitations as FID, and despite using an adapted network FVD still tends to focus more on single frame quality than on temporal coherence which is essential to evaluate videos.

IS: The Inception Distance (IS), uses an inception network that gives the probability of an image to belong to a certain class. Then, it uses the Kullback-Leibler divergence to compute a score related to the quality and diversity of the generated images. Specifically, the score is calculated to evaluate two factor: Intra-Class Similarity (high-quality images should have a strong probability of belonging to a single class) and inter-Class Diversity (generated images should belong to a variety of classes) and is calculated as

$$\text{IS} = \exp\left(\mathbb{E}_{x\sim p_g}\left[\text{KL}(p(y|x)||p(y))\right]\right), \quad (15)$$

Where $x \sim p_g$ denotes that x is sampled from the generated images distribution p_g, $p(y|x)$ is the conditional probability distribution over the classes given the image x and $p(y)$ is the marginal probability distribution over the classes, computed as $\mathbb{E}_{x\sim p_g}[p(y|x)]$.

However similarly to FID, this metric is very reliant on the inception network. It is unable to evaluate the intra-class diversity and will not work with images of classes not seen during the training of the inception network.

LMD-F: The Landmark Distance Face (LMD-F) is a metric to evaluate TH videos. LMD-F computes the average euclidean distance between the facial landmarks extracted for a real videos and those of a generated one for the same conditioning input (e.g driving audio speech). For LMD-F all of the face landmarks are used. LMD-F is calculated as

$$\text{LMD-F} = ||x_f - x'_f||_2 \quad (16)$$

where x_f and x'_f are the facial landmarks of the real and generated video respectively.

While LMD-F has been shown to correlate better than other metrics with human evaluation [55], it still suffers from being a direct comparison to the ground truth. Indeed, different head motions and expressions between the generated sequence and the ground truth will be strongly penalized. At the same time, this difference is expected since head motion and expression are only loosely correlated to the audio sequence. However, as long as both look natural, human evaluators will give a high rating to the video even if it is different from the ground truth.

LMD-M: The Landmark Distance Mouth (LMD-M) is a metrics to evaluate TH videos. LMD-M compute the average euclidean distance between the facial landmarks extracted for a real videos and those of a generated one for the same conditioning input (e.g driving audio speech). For LMD-M only the landmarks pertaining to the mouth area landmarks are used.

$$\text{LMD-M} = ||x_m - x'_m||_2 \tag{17}$$

where x_m and x'_m are the mouth landmarks of the real and generated video respectively.

While LMD-M has been shown to correlate better than other metrics with human evaluation [55], it still suffers from being a direct comparison to the ground truth. This direct comparison causes small temporal lags to be penalized when it wouldn't be noticed by human evaluators according to the recommendation by the International Telecommunication Union [1].

LPIPS: The Learned Perceptual Image Patch Similarity (LPIPS) measures the perceptual similarity between two images and try to provide a score that align with human perception. LPIPS uses a pre-trained CNN to obtain deep-features and computes the similarity between these features. The LPIPS value of CNN layer l is calculated as

$$\text{LPIPS}_l(x, x') = \sum_l w_l \cdot ||f_l(x) - f_l(x')||_2, \tag{18}$$

where $f_l(x)$ and $f_l(x')$ are the feature representations of the real image x and the generated image x' at layer l, w_l are the weights of layer l. The final LPIPS score is a weighted sum of the LPIPS_l across all the layers of the network.

While LPIPS aligns better with human evaluation it is still very dependent on the pre-trained network and is sensitive to image alignment.

PSNR: The Peak Signal-to-Noise Ratio (PSNR) compares two images at the pixel level by measuring the ratio between the maximum possible power of a signal (the original image) and the power of corrupting noise (the generated image). It is calculated as

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{19}$$

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left[I(i,j) - K(i,j) \right]^2, \tag{20}$$

where MAX is the maximum possible pixel value of the image (usually 255), $I(i,j)$ represents the pixel value at position (i,j) in the original image, and $K(i,j)$ represents the pixel value at position (i,j) in the reconstructed image. The sums are taken over all pixels in the $m \times n$ image.

PSNR does not take the structure of the image into account, is very sensitive to noise and to outliers which can lead to low correlation with human evaluation.

SSIM: The Structural Similarity (SSIM) is a score that evaluates the similarity between two images. It is obtained by combining three components : the difference in brightness between the images, the difference in contrast between the images and the structural similarities between the images across small patches. SSIM is calculated as

$$\text{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y) \tag{21}$$

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{22}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{23}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{24}$$

where μ_x and μ_y are the means of the images x and y, σ_x and σ_y are the standard deviations of the images x and y, σ_{xy} is the covariance between the images x and y, C_1, C_2, and C_3 are small constants to stabilize the division when the denominators are close to zero. $l(x, y)$, $c(x, y)$ and $s(x, y)$ correspond to the luminance comparison, contrast comparison and structure comparison respectively.

SSIM need the images to be perfectly aligned in order to be accurate. Also, since it use small patches, it focus on local structure rather than global which lead to low correlation with human perception on complex images.

Syncnet: Syncnet [8] is a CNN-based network, aims to capture the correlation between audio and spatio-temporal features of the mouth region, calculating the audio offset (the number of frames by which audio is early or late compared to video). Its distance (LSE-D) and confidence score (LSE-C) are widely use to evaluate audio-lip synchronization in TH video. While Syncnet is good at evaluating the audio offset, finding the speaker in a video containing multiple persons or detecting unrelated audio (e.g dubbing) it is less useful when comparing two videos with similar lip synchronization (e.g videos generated by two different methods). In fact it has been shown that LSE-C and LSE-D have very limited correlation with human evaluation [55]. Some recent methods [45] were even able to outperform the ground truth by a large margin on these metrics, showcasing their limitations. Additionally recent works [48, 49] have shown that Syncnet is not stable and can easily be influenced by factors outside of lip synchronization (e.g mouth cropping, image quality, brightness...) making it difficult to apply on the diverse datasets used today. Additionally, our own experiments have shown that Syncnet is sensitive to audio and video encoding even when there are not noticeable difference for a human observer (SM B)

LSE-D: The Syncnet Distance (LSE-D) compute the distance between audio and video features at the offset predicted by Syncnet. See **Syncnet** entry for limitations.

LSE-C: The Syncnet Confidence score (LSE-C) computes the difference between the minimum and the median of the features distances over all possible offsets ($-10 \leq offset \leq 10, \quad offset \in \mathbb{Z}$). See **Syncnet** entry for limitations.