Talking Face Generation by Tri-Flow: A Streaming Method

Hongwei Hu, Wangzixi Zhou, Bagus Tris Atmaja and Sakriani Sakti
Nara Institute of Science and Technology, Japan
E-mail: {hu.hongwei.hl3, zhou.wangzixi.az3, bagus.tris}@naist.ac.jp, ssakti@is.naist.jp

***Abstract*—We present a novel streaming talking face generation framework that leverages latent-space flow matching alongside a future-aware audio conditioning mechanism. By mapping high-dimensional pixel data into a compressed latent space via a pre-trained autoencoder, our model significantly reduces computational complexity. Furthermore, by conditioning the generation of each frame on both the current visual state and a sliding window of future audio slices, our approach substantially strengthens the audio signal's ability to drive and control dynamic facial movements. This enhanced cross-modal control ensures that the generated expressions are highly fluid and natural, with minimal artifacts. Extensive empirical evaluations demonstrate that our proposed method achieves a competitive Fréchet Video Distance (FVD) of 1.80 at 256×256 resolution, presenting a robust solution for high-quality, streaming-capable facial animation.**

***Index Terms*—Facial animation, flow matching, latent space, streaming generation.**

I. Introduction

Audio-driven facial animation has become a cornerstone technology for various applications, including virtual avatars, telepresence, and digital entertainment. The primary goal is to synthesize highly realistic and temporally coherent facial movements that synchronize with an input speech signal with the least sense of incongruity. While offline generation methods have achieved remarkable visual fidelity [1], streaming facial animation—where frames are generated sequentially in real-time or near real-time—remains a formidable challenge.

Frameworks such as Generative Video Bi-flow [2] have demonstrated the potential of bidirectional flow matching to maintain temporal consistency in streaming contexts. However, these foundational approaches suffer from two critical limitations. First, they operate directly in the pixel domain. As target resolutions scale up, the computational burden of pixel-space flow matching grows quickly, often leading to blurry outputs and performance degradation in visual fidelity [3]. Second, standard streaming frameworks often generate the current frame conditioned strictly on the preceding frame and the immediate audio context. This strictly causal approach severely limits the model's ability to capture how audio signals effectively drive and control overall facial expressions.

To address these critical bottlenecks, we present a novel streaming facial animation framework, Tri-flow, that elevates flow matching from the pixel domain to the latent space while introducing a future-aware audio conditioning mechanism. There are three processes in the flow matching: de-noise, video (image), and audio, so we call it Tri-flow. By mapping high-dimensional pixel data into a compressed, low-dimensional latent space using a pre-trained autoencoder, our model reduces computational complexity. This allows the flow matching process to focus entirely on learning the semantic temporal dynamics of facial expressions, enabling high-fidelity synthesis at a 256×256 resolution.

Our proposed method shows measurable improvements over existing baselines. When evaluating expression generation, the original Generative Video Bi-flow model struggles with a Fréchet Video Distance (FVD) of over 30 and degrades to an FVD of 74.07 when retrained from scratch. In contrast, our latent, future-aware approach achieves an FVD of 1.80 at 256×256 resolution, showing a substantial improvement in generation quality. In summary, our main contributions are as follows:

- **Latent Flow Matching for Video:** We adapt Bi-flow matching into the latent space for streaming facial animation, effectively solving the resolution bottleneck of pixel-space methods and enabling consistent 256×256 generation,
- **Future-Aware Audio Conditioning:** We introduce a 1-second audio lookahead buffer utilizing a robust Wav2vec encoder [4], allowing the model to an-

Part of this work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP23K21681, and JP24K02967, as well as JST NEXUS (JP- MJNX25C1).

ticipate phonemes and generate more natural facial movements,

- **Improved Performance:** We demonstrate enhanced visual quality for streaming facial animation, reducing the FVD from the baseline's $\sim$74.07 down to 1.80 (at 256×256 resolution).

II. Related Work

A. Audio-Driven Facial Animation

Traditional methods often relied on intermediate representations like 3D Morphable Models (3DMMs) [5]. While providing explicit control, they often suffer from a domain gap in photorealistic rendering [6]. Instead, our method bypasses these intermediate 3D structures entirely and operates directly on compressed 2D image latents, which effectively eliminates this rendering domain gap and allows for the direct synthesis of photorealistic, high-fidelity facial expressions.

End-to-end approaches using GANs or Diffusion Models, such as Wav2Lip [7], achieve impressive synchronization but often operate in an offline, non-streaming manner. Unlike these approaches, Tri-Flow is explicitly designed for real-time streaming, ensuring both tight audio-visual synchronization and temporal coherence without the need for full-sequence offline processing.]

B. Generative Video Bi-flow and Flow Matching

Flow Matching [8] is a simulation-free training method for neural ODEs that learns a vector field to transform a source distribution into a target distribution [9]. In the context of video, standard conditional diffusion models typically generate each frame starting from pure noise, which is computationally expensive due to the large number of ODE solver steps required [10].

A foundational advancement in this field is *Generative Video Bi-flow* [2]. The core contribution of Bi-flow lies in its joint flow formulation. Specifically, it simultaneously models a *video flow* to capture the temporal transition between frames, and a *de-noising flow* to actively correct accumulated errors. By unifying these two processes, it effectively mitigates temporal drift and eliminates the lengthy denoising process associated with starting from pure noise for each frame [11].

However, the authors only apply this method to the 128×128 pixel domain, which is computationally bottlenecked and prone to visual degradation when scaling to higher resolutions. In contrast, our Tri-flow method not only upgrades the resolution to 256×256 by using a compressed latent space, but also integrates the audio flow into the vector field, transforming it into a robust framework specifically tailored for speech-driven, streaming-style facial animation.

C. Audio Representation and Temporal Conditioning

The choice of audio feature extraction is critical for effectively driving facial dynamics. While modern frameworks predominantly utilize self-supervised representations like Wav2vec [4] or HuBERT, strictly causal streaming models often suffer from weak audio conditioning, limiting the audio signal's ability to fully control and influence the generated expressions [12]. To overcome this, our framework introduces a future-aware conditioning mechanism by buffering 1 second of future audio. By providing the generative model with this lookahead context, we strengthen the audio's control over the output, yielding highly natural and fluid expressions while maintaining a fixed latency suitable for streaming-style interaction.

III. Proposed Method

To address the limitations of pixel-space flow matching and the inadequate audio control over facial dynamics in existing streaming models, we propose *Tri-Flow*. Our framework consists of three core components: latent space projection, future-aware audio conditioning, and an identity-preserving joint flow matching mechanism. Fig 1 illustrates our framework.

A. Latent Space Projection

Standard Generative Video Bi-flow models operating on RGB pixels suffer from computational bottlenecks and visual degradation at higher resolutions (e.g., 256×256). To resolve this, Tri-Flow shifts the generative process into a compressed latent space. We utilize the pre-trained AutoencoderKL (VAE) from Stable Diffusion 3 Medium [13] to encode each high-dimensional frame into a compact latent representation $z \in R^{16\times32\times32}$ (for a 256×256 input). This spatial compression substantially reduces the computational overhead for the Diffusion Transformer (DiT) backbone, enabling the model to focus on semantic temporal dynamics of facial expressions rather than high-frequency pixel details [14].

B. Audio Conditioning

We extract dense acoustic features using a pre-trained Wav2vec encoder [4], which outperformed Whisper [15] and Qwen-TTS [16] in our evaluations. For generating a frame at time t, rather than relying solely on the aligned audio token, we utilize a 1-second lookahead window starting from t. This sequence, c_{audio}, provides

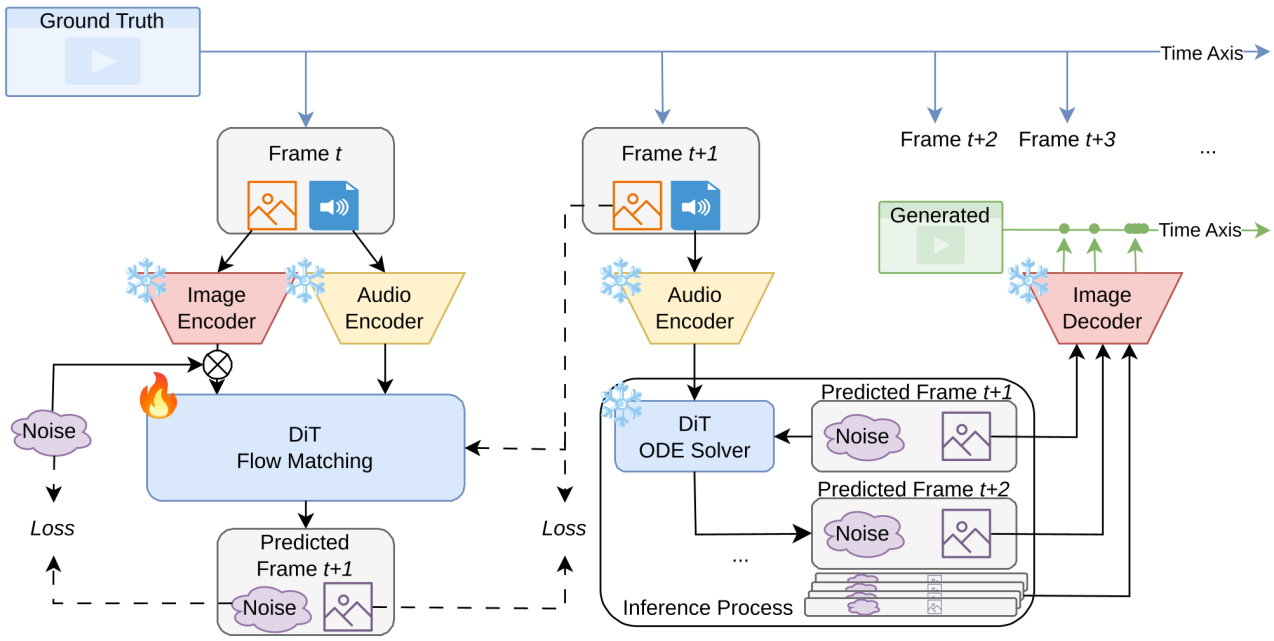


Fig. 1. Proposed Tri-Flow framework. The process begins with encoding video and audio frames into latent space, followed by joint flow matching. The Flow matching needs to match both the noise and the target image. When in inference, the process turns into an autoregressive pattern. The output image will become the next input image by subtracting the predicted noise.

immediate future phonetic context, enabling the model to anticipate mouth shapes prior to full articulation. During inference or at the sequence boundary, unavailable future audio frames are zero-padded.

C. Joint Flow Matching

We formulate the streaming generation as learning a joint vector field between consecutive latent frame distributions. To ensure the generated face maintains consistent identity attributes, we explicitly condition the DiT backbone on the initial video frame z_{ref}.

To prevent error accumulation over long sequences, we employ a bi-linear training objective. Let z_0 and z_1 represent consecutive latent chunks. We define a bi-linear interpolation state $z_{t,\alpha}$ as:

$$z_{t,\alpha} = z_0 + t(z_1 - z_0) + \alpha n, \tag{1}$$

where $t \in [0, 1]$ represents the temporal transition, α is the noise magnitude, and $n \sim \mathcal{N}(0, I)$ is Gaussian noise simulating accumulated drift.

Our DiT backbone takes the noisy state $z_{t,\alpha}$, the temporal step t, the noise level α, the future-aware audio features c_{audio}, and the anchor frame z_{ref} as inputs. The model jointly predicts two components: the temporal velocity field v_θ (predicting $z_1 - z_0$) and the denoising field n_θ (predicting n). The overall objective function is defined as:

$$\mathcal{L} = \lambda \cdot \text{MSE}(v_\theta, z_1 - z_0) + \text{MSE}(n_\theta, n), \tag{2}$$

where λ is a balancing weight (empirically set to account for chunk size variations).

During inference, we achieve stable streaming via a joint sampling ODE solver. Starting from the previously generated latent chunk with an injected noise level ϵ, the characteristic ODE is solved from $(t = 0, \alpha = \epsilon)$ to $(t = 1, \alpha = 0)$ using the Euler method. The joint velocity field governing this trajectory is formulated as:

$$\frac{dz}{dt} = v_\theta(z, t, \alpha, c_{audio}, z_{ref}) - \epsilon \cdot n_\theta(z, t, \alpha, c_{audio}, z_{ref}) \tag{3}$$

This formulation allows the system to smoothly advance to the next frame (v_θ) while simultaneously removing accumulated errors ($\epsilon \cdot n_\theta$), yielding a robust, high-fidelity facial animation stream.

IV. Experiments

A. Experiment Conditions

- **Dataset:** We evaluate on a preprocessed HDTF dataset [17], utilizing a Stable Diffusion 3 Medium VAE [13] for visual latent encoding and Wav2vec2-base-960h [4] for audio feature extraction.

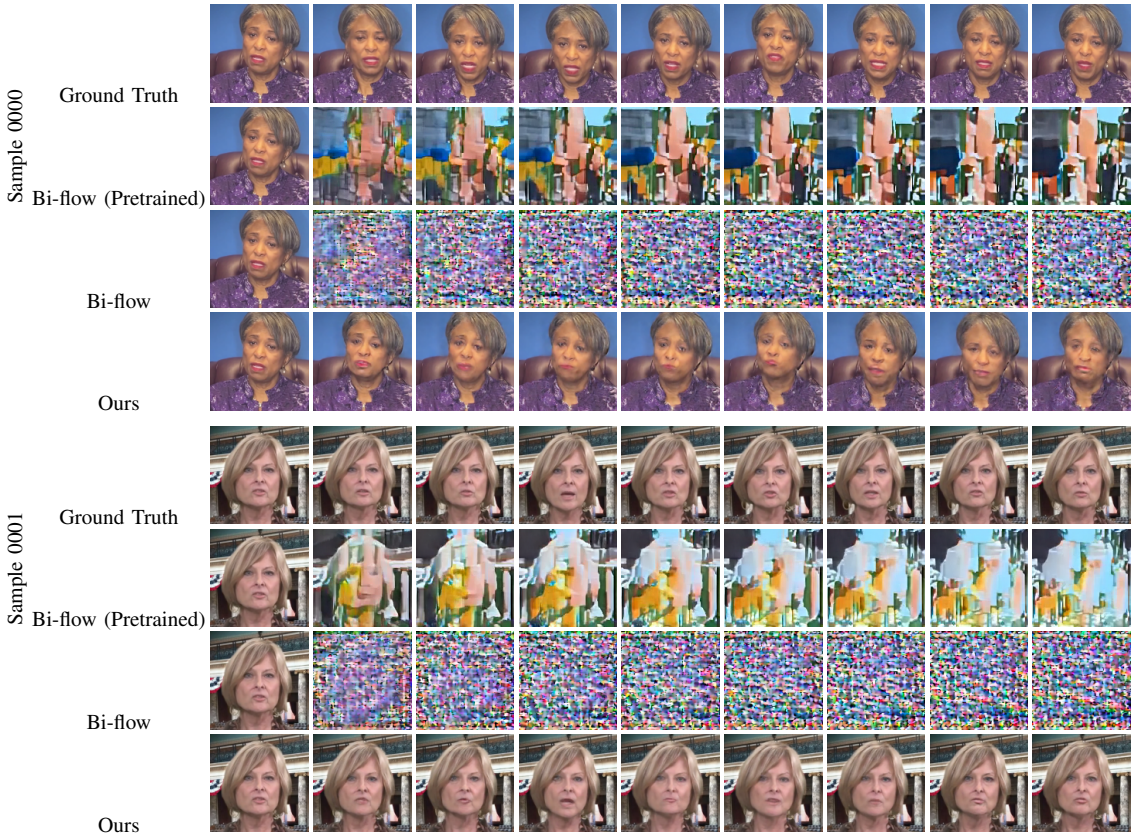


Fig. 2. Qualitative comparison of facial expression generation across different models. From top to bottom for each sample: Ground Truth, Generative Video Bi-flow (pretrained weights), Bi-flow (retrained on HDTF), and our proposed Tri-Flow method. Frames are sampled every 10 time steps.

- **Implementation Details:** Our Tri-flow approach employs three processes: video, denoise, and audio, powered by a DiT-B/4 backbone ($\sim$130M parameters). Training utilizes bf16 mixed precision, an Euler ODE solver, a batch size of 10 (gradient accumulation 8, clipping 1.5), and seed 42, following the standard HDTF split (80/10/10 for train/val/test), defaults to 20 ODE steps, with $\alpha = 0.1$. The learning rate follows a Cosine Annealing schedule from 1×10^{-4} to 1×10^{-6}. All experiments run on NVIDIA RTX PRO 6000 and H100 GPUs.
- **Baselines & Metrics:** We compare against Generative Video Bi-Flow, retraining its U-Net backbone (up to 2000 epochs) to select the lowest validation loss checkpoint. Generation quality is evaluated using Fréchet Video Distance (FVD) [18] with the pretrained I3D feature extractor [19].

B. Evaluation Results

- **Quantitative Results:** We measure the generation quality of our proposed method against the baseline Bi-flow models. As presented in Table I, applying the original pretrained Bi-flow weights to various scenarios yields an average FVD of 36.15. When retraining the Bi-flow model specifically on the HDTF dataset using pixel-space embeddings, the FVD is evaluated at 74.07 (the output is resized to 128×128 for a fair evaluation).

TABLE I
EVALUATION FOR BI-FLOW MODELS AND OUR MODELS ON THE HDTF DATASET

Model Settings	**FVD** $\downarrow$
Bi-flow (biking)	39.07
Bi-flow (mazes)	**26.59**
Bi-flow (minerl)	36.70
Bi-flow (no-traffic)	38.24
Bi-flow (riding)	44.99
Bi-flow (sky)	31.33
Avg.	36.15
Bi-flow (Retrained on HDTF, 128×128)	74.07
Tri-flow (Ours, 128×128)	**1.29**
Tri-flow (Ours, 256×256)	**1.80**

Our proposed method, which shifts the flow matching process to the latent space, achieves a sig-



Fig. 3. Generated samples with fixed audio input. The model can adapt the face to different expressions while changing the audio input.

TABLE II
ABLATION STUDY ON AUDIO ENCODERS (EVALUATED AT 128×128)

Audio Encoder	FVD ↓
Wav2vec (selected)	**1.29**
Whisper	1.55
Qwen-TTS	2.30

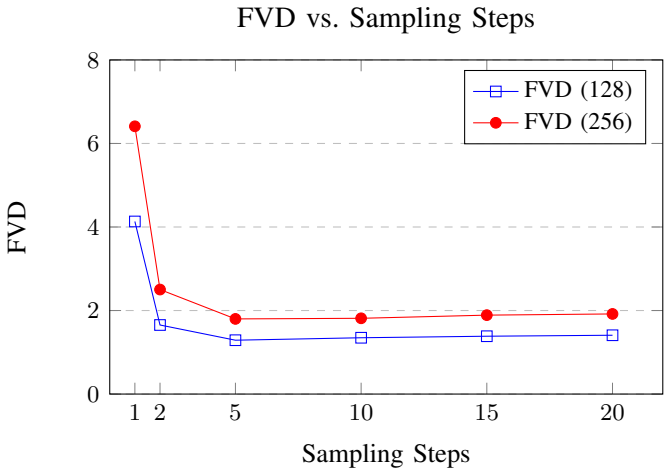


Fig. 4. Quantitative analysis of generation quality across different sampling steps. The model shows rapid convergence and stable performance beyond 5 steps in both cases.

nificantly lower FVD of 1.29 at the 128×128 resolution, and successfully scales to 256×256 with a competitive FVD of 1.80. This indicates that our latent space adaptation is highly effective in maintaining visual fidelity.

We perform a study on the audio encoders, evaluating the performance of Wav2vec against Whisper and Qwen-TTS under the 128×128 resolution (Table II). The results indicate that the Wav2vec encoder provides the most robust features for this specific generation task. In Fig. 4, we also show the FVD scores with different ODE steps in both 128×128 and 256×256 resolutions. We experiment with different checkpoints and resolutions to enhance credibility. Our model shows rapid convergence and stable performance beyond 5 steps, while Bi-flow models need approximately 15 steps on their datasets [2].

We also measured the FPS for 256×256 on two NVIDIA RTX PRO 6000 GPUs, separating DiT and VAE into different devices. Our model can achieve an average FPS of 76.52 with bf16 precision and 5 ODE steps for generating one sample with 81 frames. The proportion of time consumed by the DiT model and the VAE decoding operation is about 80% and 20% separately.

Additionally, most samples maintain an InsightFace [20] temporal cosine similarity well above 0.4 for 81 frames—a threshold already indicating high identity consistency. Notably, some cases sustain an ultra-high similarity of more than 0.75 across all 81 frames, while others hold this peak for at least 15 frames.

- **Qualitative Results:** As illustrated in Fig. 2, Tri-Flow generates high-fidelity talking-face video sequences without noticeable degradation over time. Visual comparisons with baseline methods in the same figure highlight the effectiveness of the latent space design in mitigating accumulated errors. Furthermore, Fig. 3 demonstrates that our model can adaptively adjust facial expressions based on the audio input.

V. CONCLUSION

We present Tri-Flow, a latent-space flow matching framework for high-fidelity, streaming-style talking face generation. By operating in a compressed latent space and incorporating a future-aware audio condition, our method overcomes traditional pixel-space bottlenecks to achieve stable and fast 256×256 synthesis. Achieving a

highly competitive FVD of 1.80, Tri-Flow substantially outperforms existing bi-flow baselines, offering a robust and scalable solution for real-world applications. A more extensive evaluation across baselines and lookahead lengths would be an interesting direction for future work.

REFERENCES

[1] H. K. Kim, S. Lee, and H. G. Kim, "Memorytalker: Personalized speech-driven 3d facial animation via audio-guided stylization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2025, pp. 11 241–11 251.

[2] C. Liu and T. Ritschel, "Generative video bi-flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025, pp. 19 363–19 372.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.10752

[4] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, *Wav2vec 2.0: A framework for self-supervised learning of speech representations*, 2020. arXiv: 2006.11477 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2006.11477

[5] W. Zhang et al., *Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation*, 2023. arXiv: 2211.12194 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2211.12194

[6] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, *Pirenderer: Controllable portrait image generation via semantic neural rendering*, 2021. arXiv: 2109.08379 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2109.08379

[7] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, ACM, Oct. 2020, pp. 484–492. DOI: 10.1145/3394171.3413532 [Online]. Available: http://dx.doi.org/10.1145/3394171.3413532

[8] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow matching for generative modeling*, 2023. arXiv: 2210.02747 [cs.LG].

[9] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, *Neural ordinary differential equations*, 2019. arXiv: 1806.07366 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1806.07366

[10] J. Song, C. Meng, and S. Ermon, *Denoising diffusion implicit models*, 2022. arXiv: 2010.02502 [cs.LG].

[11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, *Video diffusion models*, 2022. arXiv: 2204.03458 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.03458

[12] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, *Capture, learning, and synthesis of 3d speaking styles*, 2019. arXiv: 1905.03079 [cs.CV].

[13] P. Esser et al., *Scaling rectified flow transformers for high-resolution image synthesis*, 2024. arXiv: 2403.03206 [cs.CV].

[14] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, 2022. arXiv: 2212.04356 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2212.04356

[16] H. Hu et al., *Qwen3-tts technical report*, 2026. arXiv: 2601.15621 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2601.15621

[17] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3660–3669. DOI: 10.1109/CVPR46437.2021.00366

[18] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, *Towards accurate generative models of video: A new metric challenges*, 2019. arXiv: 1812.01717 [cs.CV].

[19] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[20] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *CoRR*, vol. abs/1801.07698, 2018. arXiv: 1801.07698.