Disentangled Motion Diffusion for Long-Form Voice-Driven Portrait Animation

Yi-Chun Chang
Arete Honors Program
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
george930502.sc11@nycu.edu.tw

Po-Hsuan Fan Chiang Jen-Tzung Chien
Institute of Electrical and Computer Engineering
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
{ed223.ee13, jtchien}@nycu.edu.tw

***Abstract*—Long-form portrait animation with high-fidelity and low-latency has been highly-demanding when building the human-computer interactive systems. The existing methods often suffer from inherent trade-off between effectiveness and efficiency while improving one objective but compromising another. This paper presents a novel audio-driven portrait animation, which mitigates the compromise through enhancing the representation by flow-divergence matching in sequence animation under an orthogonal latent space for motions. To strengthen the facial motion controllability, the latent space is decomposed into a transferable motion subspace for identity-agnostic dynamics and a preservable subspace for portrait-specific traits. Furthermore, a sparse motion dictionary is constructed to fulfill the fine-grained disentanglement, allowing an independent control over semantic motion factors and improving both realism and identity consistencies in portrait animation. To speed up the generation, we employ a transformer-based vector field predictor that leverages flow-based temporal fusion to integrate intra-clip and inter-clip audio perception, resulting in temporally-coherent and contextually-aware facial dynamics. Extensive experiments demonstrate that our method achieves stable long-term video synthesis and outperforms state-of-the-art baselines in terms of video quality, generation speed, and temporal coherence.**

***Index Terms*—orthogonal latent bases, flow and divergence matching, flow-based temporal fusion, talking portrait**

I. Introduction

With the growing demand in human-computer interaction applications, portrait animation has attracted increasing attention. Beyond interactive systems, it plays a key role in the applications such as video conferencing, virtual avatars, and e-education. Portrait animation aims to synthesize realistic and temporally coherent facial motion from static images, typically conditioned on an external input such as audio signal [1], motion trajectory [2] or semantic guidance [3]. Despite its promising applications, voice-driven portrait animation of a previously unseen identity remains highly challenging, as it requires faithfully capturing subtle yet critical facial cues from speech, including precise lip-audio synchronization, natural eye gaze and blinking, realistic head movements, and variations in facial behavior across different expressions. Along with generation fidelity, practical concerns for real-time response and stable long-form generation are also essential in interactive and communication scenarios [4], [5]. As illustrated in Fig. 1, our method addresses these challenges by generating natural, low-latency talking portraits from a single reference image and a long-form audio stream.

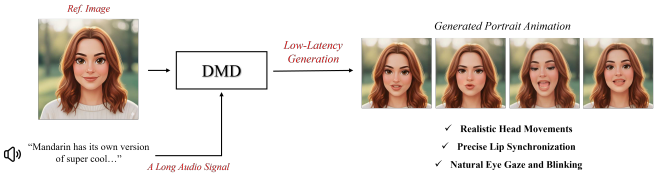

Fig. 1. Given a reference portrait and a long-form audio stream, our method animates a natural low-latency talking head by enhancing motion representation through flow-divergence matching under orthogonal latent spaces containing natural head motion, precise lip synchronization, and authentic eye gaze and blinking.

However, existing approaches suffer from inherent trade-offs among quality, speed, and consistency. Diffusion-based models [3], [6] achieve realistic animation but require heavy computation. Real-time methods [7] reduce latency but sacrifice fidelity in fine-grained facial details and disentanglement. 3D Gaussian splatting [8] methods [9], [10] deliver comparable quality with improved speed, but are constrained to single-identity training and inference, limiting the cross-identity generalization.

Self-supervised methods using orthogonal motion codes [11], [12] address this by learning interpretable motion semantics with minimal supervision. Riemannian geometric analyses [13] further show that navigating latent bases at selected time steps enhances semantic control. Meanwhile, flow matching [14], [15] has emerged as an efficient alternative to diffusion, enabling faster sampling with improved generalization.

This paper presents a voice-driven portrait animation framework that integrates several complementary techniques into a unified two-stage pipeline. Our formulation combines flow-divergence matching [15] with an orthogonal motion latent space [11], [16], where the latent space is decomposed into preservable and transferable subspaces [12] to disentangle identity-preserving attributes from motion dynamics, enabling generalization across different identities. To further enhance

This research was supported by the National Science and Technology Council (NSTC) under the College Student Research Scholarship (Grant No. 113-2813-C-A49-020-E).

interpretability, we incorporate a sparse motion dictionary [17] with L1-regularized coefficients, which encourages each motion instance to activate only a minimal subset of semantic bases. For fast and generalized generation, a transformer-based vector field predictor [18] conditioned on multi-scale audio features [1] is trained via the flow-divergence matching objective [15], which jointly aligns the vector field and its divergence to reduce the residual error between learned and ground-truth conditional flows. While individual components draw from prior work, the principal contribution lies in their integration and the resulting synergy: the orthogonal latent space provides structured motion targets for flow matching, the sparse dictionary enhances disentanglement within this space, and the temporal fusion mechanism exploits the flow trajectory to maintain long-form coherence. Experimental results demonstrate that this combination enables stable and sustained frame synthesis, achieving high-fidelity and low-latency talking portraits. Our contributions are summarized as follows:

- Integrating flow-divergence matching with an orthogonal motion latent space, decomposed into preservable and transferable subspaces with L1-regularized sparse motion dictionary, to achieve disentangled, interpretable, and identity-generalized portrait animation.
- Designing a transformer-based vector field predictor with flow-based temporal fusion that continuously shifts latent positions along the flow trajectory, integrating intra-clip and inter-clip audio perception for temporally-coherent long-form generation.
- Demonstrating through extensive experiments on HDTF and CelebV-HQ that the proposed framework outperforms state-of-the-art baselines in video quality, generation speed, and temporal coherence, while generalizing to virtual faces.

Our training and inference code are publicly available.[1]

II. Related Works

This paper animates a talking portrait with a transferrable identity by editing latent space based on flow matching.

A. Latent Space Editing

Exploring semantically meaningful representation in latent space has been a central research topic in machine learning [19], as it enables interpretable representation learning and intuitive image manipulation. With the advances in image generation, extensive efforts have been devoted to semantic face editing by navigating the latent space via generative adversarial network (GAN). Both supervised [20] and unsupervised [21] methods have been developed to modify facial attributes through the controlled shifts in latent distributions while maintaining photo-realistic outputs. Recent advances have introduced a linear motion decomposition and a linear latent navigation as the core schemes to construct interpretable representation of facial dynamics [11], [12]. These methods learn a set of orthogonal bases and express the motion as linear combinations of the bases, enabling the disentangled [22], controllable, and theoretically grounded facial animation. Furthermore, Riemannian geometric analyses of diffusion model latent spaces [13] have revealed that navigating along specific basis vectors at the selected time steps could enhance the semantic control and the editing precision.

B. Flow Matching

Flow matching [14] has emerged as a competitive alternative to diffusion-based generative model [23], offering a better trade-off between computational cost and sample quality. Unlike conventional diffusion models, which rely on stochastic denoising process, flow matching formulates the generation as a deterministic transformation by regressing a vector field that interpolates between a prior noise distribution and the target data distribution. The generative process is obtained by solving the corresponding ordinary differential equations [24].

A key advantage of flow matching lies in its flexibility to design suitable conditional probability paths and vector fields, enabling more efficient sampling. One widely used design was based on the optimal transport displacement interpolation [14], which adjusted the data distribution along straight-line trajectories at constant speed. The study in [15] pointed out that while conditional flow matching efficiently approximates the probability paths, its divergence gap can cause significant errors in path learning and likelihood estimation. To address this issue, an upper bound on the gap was bridged by minimizing a conditional divergence loss, yielding a new training approach to flow and divergence matching, that significantly improved the flow matching across challenging tasks.

C. Portrait Animating

Prior works on voice-driven portrait animation are grouped into four categories. First, the facial landmark and 3D morphable models [25] explicitly represented the facial geometry, constrained by the limited expressiveness of 3D meshes, which restricted the fine-grained detail and the natural motion. Second, the 2D GAN-based methods [26]–[28] directly synthesized the portraits in image space but suffered from the temporal instability, inaccurate lip-audio synchronization, and degraded quality due to long sequences. Third, the diffusion-based models [2] enhanced the robustness and motion diversity through probabilistic modeling, yet their large-scale networks and iterative sampling hindered the real-time performance. Fourth, the 3D Gaussian splatting-based methods [9], [10] achieved high-fidelity 3D-consistent rendering and novel-view synthesis but remained limited to single-identity training and exhibited weak cross-identity generalization.

Recent methods [16], [29] have attracted significant attention by achieving real-time talking portrait generation via diffusion transformer [18], [30] and flow matching [14] in the latent space of motion movements. Our work shares the latent flow matching paradigm with Float [16], but differs in three aspects: (1) we adopt flow-divergence matching [15] rather

[1] https://github.com/George930502/Disentangled-Motion-Diffusion-for-Long-Form-Voice-Driven-Portrait-Animation

than standard flow matching to reduce the gap between learned and true probability paths, (2) we decompose the motion space into preservable and transferable subspaces with L1-regularized sparse activation for enhanced disentanglement, and (3) we employ a position-shift temporal fusion mechanism for long-form coherence.

III. Preliminary

This section reviews the two foundational techniques upon which our framework is built: (1) the linear latent motion representation that provides the structured motion space, and (2) flow-based generative models that enable efficient generation within this space. Understanding both is essential because our method operates at their intersection, where orthogonal motion bases serve as structured targets for flow-divergence matching.

A. Linear Latent Motion Representation

Latent image animator [11] learns an autoencoder framework based on a motion encoder E_m and a generator G. This model addresses motion transfer through linear transformation, mapping from a source image $\mathbf{x}_s$ to a driving image $\mathbf{x}_d$. Within the latent space, this method learns a latent trajectory that connects the source representation $\mathbf{z}_{s\to r}$ to the driving representation $\mathbf{z}_{s\to d}$ through a linear navigation path $\mathbf{z}_{r\to d}$

$$\mathbf{z}_{s\to d} = \mathbf{z}_{s\to r} + \mathbf{z}_{r\to d}. \tag{1}$$

Specifically, the motion encoder E_m encodes the source image $\mathbf{x}_s$ into a latent code $\mathbf{z}_{s\to r}$, and the driving image $\mathbf{x}_d$ in terms of motion bases via weights $A_{r\to d} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$, where each weight α_i indicates the activation strength of a corresponding motion basis $\mathbf{d}_i$. This model further learns a trainable orthogonal motion dictionary $D_m = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\}$, where each basis vector represents a fundamental motion component such as head rotation, mouth movement, or eye blinking. Under this formulation, the latent motion path $\mathbf{z}_{r\to d}$ can be represented as a linear combination of these basis vectors

$$\mathbf{z}_{r\to d} = \sum_{i=1}^{M} \alpha_i \mathbf{d}_i, \quad \text{s.t. } \mathbf{d}_i^\top \mathbf{d}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{2}$$

The orthogonality constraint between two different bases $\mathbf{d}_i$ and $\mathbf{d}_j$ ensures the disentangled motion representation, allowing complex visual transformations to be modeled as the composition of independent motion factors. Finally, the generator G decodes latent variable $\mathbf{z}_{s\to d}$ to a dense optical flow field, which warps the source frame $\mathbf{x}_s$ into the target driving frame $\mathbf{x}_d$.

B. Flow-Based Generative Models

1) Flow Matching: Flow matching (FM) [14] provides an efficient approach to learn the continuous probability flow paths between a prior noise distribution $p_0 = p$ and the data distribution $p_1 = q$. A time-dependent vector field $\boldsymbol{u}_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ defines a flow ψ_t [31] governed by the ordinary differential equation

$$\frac{d}{dt}\psi_t(\mathbf{x}) = \boldsymbol{u}_t(\psi_t(\mathbf{x})), \quad \text{with } \psi_0(\mathbf{x}) = \mathbf{x} \tag{3}$$

which implicitly transports the samples along the path $p_t(\mathbf{x}) = p_0(\psi_t^{-1}(\mathbf{x}))\det\left[\frac{\partial \psi_t^{-1}}{\partial \mathbf{x}}(\mathbf{x})\right]$, mapping from noise distribution to data distribution as t evolves from 0 to 1. Given a data sample $\mathbf{x}_1 \sim p_1(\mathbf{x}_1) \approx q(\mathbf{x}_1)$, FM defines a *conditional probability path* $p_t(\mathbf{x}|\mathbf{x}_1)$ satisfying $p_0(\mathbf{x}|\mathbf{x}_1) = p(\mathbf{x})$ and $p_1(\mathbf{x}|\mathbf{x}_1) \approx \delta(\mathbf{x} - \mathbf{x}_1)$. The corresponding conditional vector field $\boldsymbol{u}_t(\mathbf{x}|\mathbf{x}_1)$ is learned by regressing a neural network parameterized vector field $\boldsymbol{v}_t(\mathbf{x};\theta)$ with parameter θ according to the *conditional flow matching* objective

$$\mathcal{L}_f(\theta) \triangleq \mathbb{E}_{t, q(\mathbf{x}_1), p_t(\mathbf{x}|\mathbf{x}_1)}\left[\|\boldsymbol{v}_t(\mathbf{x};\theta) - \boldsymbol{u}_t(\mathbf{x}|\mathbf{x}_1)\|^2\right] \tag{4}$$

where $t \sim \mathcal{U}[0,1]$, $\mathbf{x}_1 \sim q(\mathbf{x}_1)$ and $\mathbf{x} \sim p_t(\mathbf{x}|\mathbf{x}_1)$. Our framework adopts the *optimal transport* (OT) conditional path, which connects $\mathbf{x}_0 \sim p$ and $\mathbf{x}_1 \sim q$ along a straight line at constant velocity:

$$\psi_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1. \tag{5}$$

This defines an affine Gaussian path $p_t(\mathbf{x}|\mathbf{x}_1) = \mathcal{N}(\mathbf{x}|t\mathbf{x}_1, (1-t)^2\mathbf{I})$ with target $\boldsymbol{u}_t(\mathbf{x}|\mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0$, which is adopted in our framework due to its computational efficiency and analytical tractability.

2) Flow and Divergence Matching: Minimizing $\mathcal{L}_f$ alone does not fully constrain how the learned flow connects the prior and data distributions—the model can reach a biased equilibrium where trajectories deviate in divergence. The *flow and divergence matching* (FDM) [15] addresses this by introducing a complementary *conditional divergence matching* loss:

$$\mathcal{L}_c(\theta) \triangleq \mathbb{E}_{t,\, p_t(\mathbf{x}|\mathbf{x}_1),\, p(\mathbf{x}_1)}\Big[\Big|\big(\nabla\cdot\boldsymbol{u}_t(\mathbf{x}|\mathbf{x}_1) - \nabla\cdot\boldsymbol{v}_t(\mathbf{x};\theta)\big) + \big(\boldsymbol{u}_t(\mathbf{x}|\mathbf{x}_1) - \boldsymbol{v}_t(\mathbf{x};\theta)\big)\cdot\nabla\log p_t(\mathbf{x}|\mathbf{x}_1)\Big|\Big]. \tag{6}$$

The total objective $\lambda_f\mathcal{L}_f + \lambda_c\mathcal{L}_c$ jointly aligns the vector field and its divergence, reducing the residual error between learned and ground-truth flows. We adopt FDM because motion latent distributions are sensitive to trajectory accuracy; small divergence gaps accumulate over long sequences and cause visible motion drift.

IV. Methodology

A. Latent Disentangled Motion Space

Given a colored source portrait image $S \in \mathbb{R}^{3\times H\times W}$ with height H and width W and a driving audio speech $\mathbf{a}^{1:L} \in \mathbb{R}^{L\times n_a}$ of L frames in a clip with feature dimension n_a, our goal is to generate or estimate a portrait animation

$$\widehat{D}^{1:L} = \{\widehat{D}^t\}_{t=1}^{L} \in \mathbb{R}^{L\times 3\times H\times W}. \tag{7}$$

It is crucial to control distinct motion semantics to carry out the disentangled, interpretable and controllable facial animation. This study builds a new latent image animator to characterize motions and identities as the separate latent variables. Although the previous animator in [11] built the motion dictionary, which exhibited partial semantic interpretability, their work still suffer from the mixing of multiple semantic attributes, limiting its effectiveness to control the editing of image and video generation.

1) Sparse and Orthogonal Motion Representation: To enhance the disentanglement [32], this paper introduces a *sparse motion dictionary* [17] and learns a set of orthogonal motion bases $D_m = \{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\}$, where each basis vector $\mathbf{d}_i \in \mathbb{R}^{n_d}$ corresponds to a latent motion attribute. We enforce the orthogonality $\mathbf{d}_i^\top \mathbf{d}_j = 0$ for $i \neq j$ via QR decomposition of a learnable weight matrix, which guarantees orthonormality by construction. To encourage the model to activate only a minimal subset of bases per motion instance, we impose an L1 sparsity penalty on the motion coefficients: $\mathcal{L}_{\text{sparse}} = \lambda_s \|\boldsymbol{\alpha}\|_1$, where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$ are the activation weights produced by a fully-connected head atop the motion encoder. This L1 regularization [17] drives most coefficients toward zero, so that each motion instance is reconstructed from a compact subset of interpretable bases, enabling explicit and controllable motion editing.

2) Preservable and Transferable Motion Bases: This paper further decomposes the latent motion space into two subspaces consisting of the *preservable motion bases* for source portrait informatics as well as the *transferable motion bases* for target-agnostic dynamics

$$\mathbf{z}_{s\to d} = \mathbf{z}_{s\to r} + \mathbf{z}_{r\to d} = \sum_{i=1}^{K} \beta_i \mathbf{d}_i^k + \sum_{i=1}^{M} \alpha_i \mathbf{d}_i^m \tag{8}$$

where $\{\mathbf{d}_i^k\}_{i=1}^K$ are identity preserving bases, $\{\mathbf{d}_i^m\}_{j=1}^M$ are motion moving bases. Both sets are orthonormal, i.e., $(\mathbf{d}_i^k)^\top \mathbf{d}_j^k = (\mathbf{d}_i^m)^\top \mathbf{d}_j^m = \delta_{ij}$, and cross-orthogonal by construction since all $M{=}20$ bases are obtained from a single QR factorization. We set $K{=}8$ and $M{-}K{=}12$ following the convention established by LIA [11], where $M{=}20$ bases have been shown sufficient to span the primary modes of facial motion including rigid head pose ($\sim$3 DOF), jaw and lip articulation, eye gaze, and brow movement. The orthogonality constraint ensures that these M bases form a maximally efficient covering of the motion manifold without redundancy. This decomposition ensures that identity-agnostic motions such as lip movement and head rotation generalize across identities, while subject-specific traits such as facial structure and habitual expressions are preserved.

B. Flow-Based Temporal Fusion

Unlike most existing approaches that rely on motion or visual frames for spatial-temporal generation, we adopt an audio-only conditioning scheme [1], [33] that constructs representations directly from speech, enabling both fine-grained articulation and long-range temporal coherence.

Specifically, a pretrained speech encoder [36] extracts multi-scale features from five hierarchical stages, which are concatenated and projected through three linear layers, then combined with t and $\mathbf{z}_{s\to r}$ before being fed into the vector field predictor (Fig. 3). To address the issues of limited receptive field and temporal inconsistency commonly observed in voice-driven generation methods, we propose a *flow-based temporal fusion* mechanism conditioned on a source image and an audio clip. Unlike standard autoregressive conditioning, which concatenates previous frames as context and is prone to error accumulation, our approach integrates intra-clip and inter-clip audio perception by continuously shifting latent positions along the flow trajectory [1]. Concretely, at each ODE integration step $k \in \{0, \ldots, N_{\text{fe}}{-}1\}$, the latent sequence is partitioned into clips of $T{=}25$ frames, and the clip window is cyclically shifted by an accumulated offset $\widetilde{o} = k \cdot o$ (with $o{=}7$). Each clip is processed independently by the vector field predictor, and the resulting velocity updates are scattered back to the full sequence via modular indexing. This ensures that each ODE step observes different temporal neighborhoods, enabling the model to accumulate long-range dependencies across the entire sequence without additional computational cost. To handle boundary conditions where the shifted index exceeds the sequence length, a cyclic padding strategy wraps around and reuses the initial latent variables and audio prompts. This design captures both short-term phoneme-level dynamics and long-term temporal coherence without increasing inference cost, reducing motion jitter and achieving natural alignment between speech rhythm and facial motion.

C. Overall Generation Framework

The training procedure follows a two-stage paradigm as displayed in Figure 2. In the first stage, this procedure learns a sparse motion dictionary D_m, together with the motion encoder E_{θ_m}, magnitude vector predictor F_{θ_c}, and generator G_{θ_g}. The motion encoder is a ResBlock-based network that maps an input face to a 512-dimensional appearance embedding, from which separate fully-connected heads produce the preservable coefficients $\boldsymbol{\beta} \in \mathbb{R}^K$ and transferable coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{M-K}$. The generator follows a StyleGAN2-based synthesis architecture [37]: a constant $4{\times}4$ feature map is progressively upsampled through style-modulated convolutions to $512{\times}512$ resolution, with each layer conditioned on the motion latent via adaptive style injection. A flow warping module predicts a 2D displacement field and a blending mask from the modulated features, which warp the source appearance and blend it with the synthesized output. The first-stage parameters $\Theta^1 = \{\theta_m, \theta_c, \theta_g\}$ jointly construct a disentangled motion latent space that provides structured targets for the second stage [11], [38]. The training objective combines: L1 reconstruction loss, VGG-19 perceptual loss [39], a *component perceptual loss* applied to ROI-aligned [40] crops of the left eye, right eye, and mouth, a multi-scale discriminator with non-saturating adversarial loss, three region-specific component discriminators, a Gram-matrix feature style loss [41] for texture fidelity, an ArcFace [42] identity-preserving loss, and the L1 sparsity penalty on $\boldsymbol{\alpha}$.

In the second stage, all first-stage parameters are frozen, and we train the audio encoder E_{θ_a}, feature embedding F_{θ_e}, and vector field predictor V_{θ_v} to estimate a conditional flow vector field driven by speech. The audio encoder uses a frozen Whisper-tiny [36] model to extract multi-scale features: hidden states from all four encoder layers plus the embedding layer (five scales, 384-dim each) are concatenated and projected

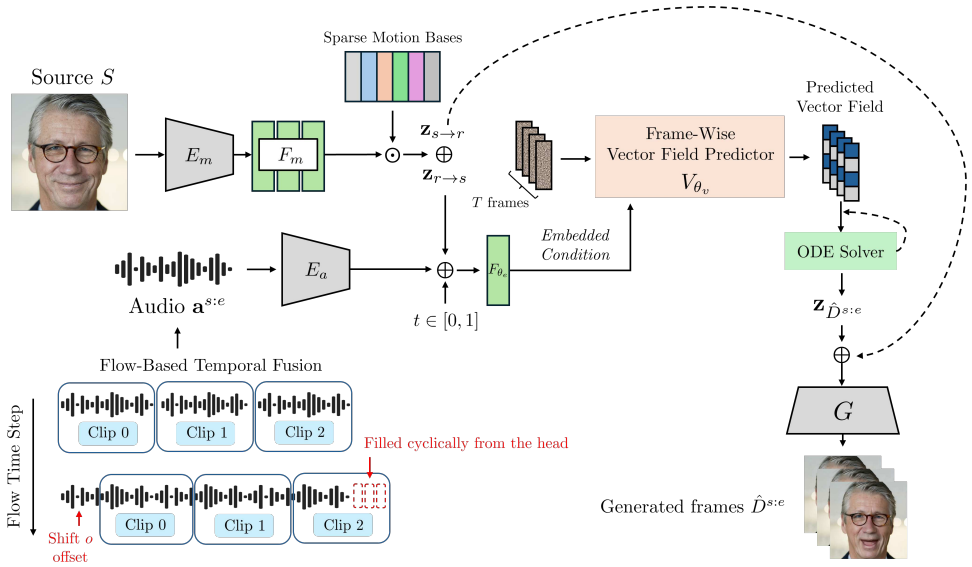

Fig. 2. Overview of the proposed framework for talking portrait animation. Given a source image S and an audio chunk $\mathbf{a}^{s:e}$, which s and e denote the start and end indices, this framework first encodes S into a latent identity $\mathbf{z}_{s\to r}$ and a latent motion $\mathbf{z}_{r\to s}$ represented by sparse identity bases and motion bases $\mathbf{d}_i^k$ and $\mathbf{d}_i^m$, respectively. Audio embeddings $\mathbf{c}_a$ are extracted to perform flow-based temporal fusion, progressively shifting the latent position offset o to capture long-term dependencies. At each flow time step $t \in [0,1]$, a frame-wise vector field predictor estimates the vector field $\boldsymbol{v}_t$ for flow and divergence matching integrated by an ODE solver to calculate the latent motion, which are combined with $\mathbf{z}_{s\to r}$ and decoded by the generator to produce the temporally coherent portrait animation $\hat{D}^{s:e}$.

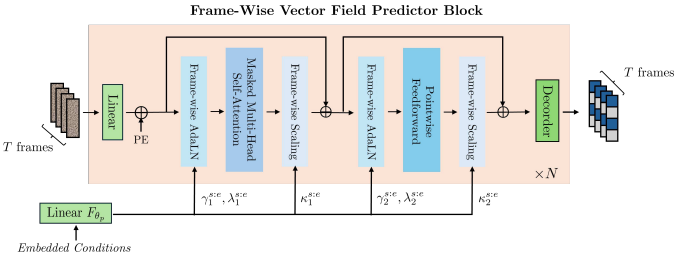

Fig. 3. Frame-wise vector field prediction. The predictor applies frame-wise AdaLN [18], masked self-attention [34], [35], and pointwise feedforward layers. Conditions are formed by concatenating encoded audio with flow time step $t \in [0,1]$ and latent identity $\mathbf{z}_{s\to r}$, projected to six AdaLN parameters $\gamma_i^{s:e}$, $\lambda_i^{s:e}$, $\kappa_i^{s:e}$ ($i \in \{1,2\}$). $\kappa_{i,t}$ serves as a frame-wise residual gate.

through a three-layer MLP (19,200→1,024→512→512) to produce per-frame audio embeddings $\mathbf{c}_a \in \mathbb{R}^{L\times 512}$. The motion latent targets $\mathbf{z}_{D^{1:L}}$ are obtained by encoding each ground-truth driving frame through the frozen first-stage encoder and computing $\mathbf{z}_{r\to d} = \boldsymbol{\alpha} \cdot D_m^\top$, yielding a sequence of 512-dimensional motion vectors. The learning objective follows the OT conditional flow matching formulation with target $\boldsymbol{u}_t(\mathbf{z}|\mathbf{z_1}) = \mathbf{z}_{D^{1:L}} - \mathbf{z}_0$, where $\mathbf{z}_0 \sim \mathcal{N}(0^{1:L}, \mathbf{I})$, linearly interpolating between noise and the structured motion representation. Additionally, we minimize the divergence discrepancy term using a Hutchinson trace estimator with Jacobian-vector products [15], computed in full precision to ensure numerical stability of the second-order gradients. This study incorporates a velocity loss to further reinforce temporal consistencies and ensure smooth motion transitions across consecutive frames according to

$$\mathcal{L}_v(\theta_e,\theta_v) \triangleq \mathbb{E}_{\mathbf{z}}\left[\|\Delta\boldsymbol{v}_t(\mathbf{z};\theta_e,\theta_v) - \Delta\boldsymbol{u}_t(\mathbf{z}|\mathbf{z_1})\|\right] \tag{9}$$

where $\Delta\boldsymbol{v}_t$ and $\Delta\boldsymbol{u}_t$ are the one-frame difference along the time-axis between the prediction $\boldsymbol{v}_t$ and the target $\boldsymbol{u}_t$ vector field. The total objective is accordingly expressed as

$$\mathcal{L}(\Theta^2) = \lambda_f\mathcal{L}_f(\Theta^2) + \lambda_c\mathcal{L}_c(\Theta^2) + \lambda_v\mathcal{L}_v(\Theta^2) \tag{10}$$

where $\Lambda = \{\lambda_f, \lambda_c, \lambda_v\}$ is the set of hyperparameters, and $\Theta^2 = \{\theta_e, \theta_v\}$ denotes the second-stage trainable parameters. The whole model contains two parameter sets $\Theta = \{\Theta^1, \Theta^2\}$ for two stages of training. Once the vector field is learned, the corresponding latent motions are obtained by solving the associated ODE, enabling temporally coherent motion generation guided by the audio sequence.

V. Experiments

A. Experimental Setup

We train on two open-source datasets: HDTF [43] (∼16 hours, 400+ identities) and CelebV-HQ [44] (35K clips, 15K+ identities). Following [26], facial regions are localized via face alignment [46], bounding boxes are temporally refined with IoU tracking, and each crop is normalized to 512×512.

1) Implementation Details: The latent motion space has dimension $d = 512$ with $M = 20$ orthogonal bases (K=8 preservable, 12 transferable). Phase 1 trains for 800K steps with Adam [47] (β_1=0, β_2=0.99, lr= 2×10^{-4}, batch size 8) and employs lazy R1 gradient penalty every 16 discriminator steps. Phase 2 trains the vector field predictor (N=8 blocks, 8 heads, h=1024, MLP ratio 4×, attention window ±2 frames) with Adam (β_1=0.9, β_2=0.999, lr= 10^{-5}, batch size 8)

TABLE I
QUANTITATIVE COMPARISON ON THE COMBINED HDTF [43] AND CELEBV-HQ [44] TEST SET. THE BEST RESULT FOR EACH METRIC IS IN **BOLD**.

Method	FID ↓	FVD ↓	CSIM ↑	E-FID ↓	P-FID ↓
SadTalker [28]	146.473 / 73.268	392.671 / 345.802	0.689 / 0.654	2.587 / 2.110	1.805 / 1.636
EchoMimic [2]	84.578 / 38.637	324.736 / 302.932	0.767 / 0.732	1.530 / 1.376	0.072 / 0.056
Hallo [45]	61.728 / 31.392	243.820 / 197.186	0.829 / 0.854	1.432 / 1.137	0.126 / 0.092
Hallo2 [3]	42.187 / 27.810	193.196 / 163.421	0.837 / 0.861	1.249 / 1.099	0.082 / 0.066
Ours	**27.829 / 22.613**	**163.110 / 132.915**	**0.862 / 0.894**	**1.122 / 0.993**	**0.043 / 0.035**

for 2,000K steps on a single NVIDIA A100 GPU. Clip length is T=25 frames with shift offset o=7. Loss weights: $\lambda_f = \lambda_v = 1$, $\lambda_c = 0.1$. We adopt the Euler method with N_{fe}=10 steps to solve the ODE during both training and inference. During training, dropout rate 0.1 is applied to $\mathbf{z}_{r\to s}$ and $\mathbf{a}^{s:e}$ for classifier-free guidance [48].

2) Evaluation Metrics: We report FID [49] and 16-frame FVD [50] for visual fidelity, CSIM [51] for identity preservation, E-FID [52] for expression accuracy, and P-FID for pose consistency. Baselines include SadTalker [28], EchoMimic [2], Hallo [45], and Hallo2 [3], all using publicly available pre-trained weights.

Fig. 4. Qualitative comparison over various methods in presence of an example in HDTF [43].

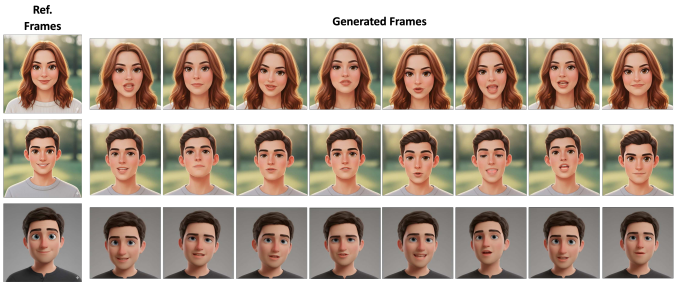

Fig. 5. Extended results using the proposed method, demonstrating a strong generalization given by the animated virtual faces (generated by Gemini) based on the trained models from real face images and videos.

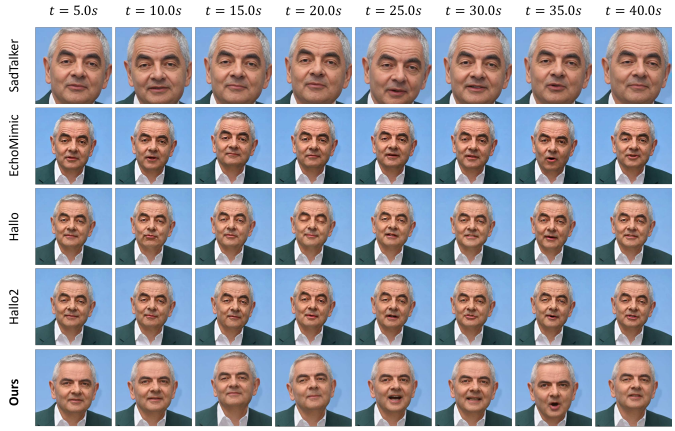

Fig. 6. Qualitative comparison over various methods by feeding a long speech input. Under the long-form audio driving, relative to other methods, our method produces high-fidelity video in lip synchronization and facial expression with natural pose variations.

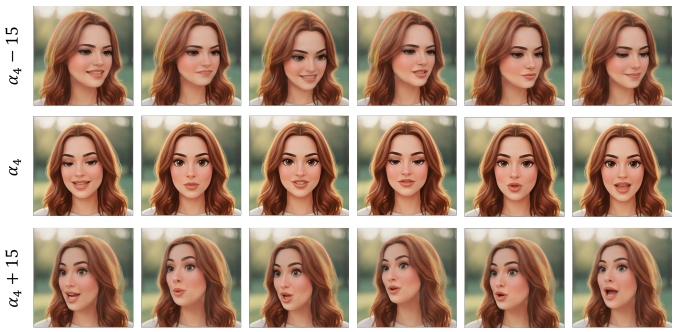

Fig. 7. Pose editing via learned orthogonal motion bases. Our model learns a 3D-aware, disentangled motion representation that enables fine-grained pose control. By manipulating the weight α_4, the identity naturally rotates left or right while preserving expression and realism, demonstrating well-structured disentanglement and semantic consistency in the learned motion space.

B. Results and Analyses

1) Qualitative Comparison: As shown in Figs. 4 and 5, our results exhibit natural head rotations and subtle micro-expressions (e.g., eye blinking and eyebrow movement) that competing methods fail to produce, instead generating over-smoothed or repetitive motion patterns. By modeling dynamics in a semantically meaningful latent space rather than relying on explicit motion frames, our framework achieves richer expressiveness, smoother pose transitions, and consistent lip synchronization even in long-duration sequences.

2) Quantitative Comparison: As shown in Table I, the proposed model consistently outperforms the related baselines on a hybrid dataset consisting of HDTF [43] and CelebV-HQ [44] benchmarks across all metrics. Our method achieves significantly lower FID and FVD scores, indicating superior image and video generation quality. The higher CSIM score reflects better identity preservation, while the reduced E-FID and P-FID values demonstrate that the generated expressions and head motions are both accurate and temporally stable. These results validate the effectiveness of our disentangled motion representation and audio-driven flow modeling in achieving realistic high-quality talking portrait generation.

3) Long-form Generation with Expressiveness: As shown in Fig. 6, our approach maintains consistent head pose diversity, facial expressiveness, and lip synchronization throughout extended speech segments. The disentangled motion space captures natural eye gaze variations and micro-expressions, while the flow-based temporal fusion enforces smooth transitions even under acoustically complex patterns.

4) Efficiency and Generalization to Virtual Faces: As shown in Table II, our model achieves a 25x-300x speedup over prior baselines, owing to its compact latent generative architecture and efficient vector field predictor. We note that the compared baselines are all diffusion-based methods operating in pixel or high-dimensional latent space, which represent the state-of-the-art quality tier; lightweight non-diffusion methods [7] sacrifice fidelity for speed and thus occupy a different design regime. Unlike diffusion-based approaches requiring iterative denoising, flow matching formulates generation as a continuous transformation guided by an estimated velocity field, requiring only $N_{\text{fe}}{=}10$ Euler steps. Beyond efficiency, our model generalizes effectively to virtual and synthetic avatars, demonstrating strong cross-identity and cross-domain robustness. As illustrated in Figure 7, the learned orthogonal motion bases support fine-grained pose editing. Adjusting the latent coefficient α_4 produces smooth and semantically consistent head rotations, validating the model's disentangled and 3D-aware motion representation.

5) Discussion of Design Choices: We further justify each key design decision based on prior evidence and our experimental observations.

(1) *Preservable vs. transferable decomposition*: ViCo-Face [12] demonstrated that splitting bases into identity-preserving and motion-transferring subsets significantly improves cross-identity reenactment quality; our CSIM scores (Table I) confirm this benefit.

(2) *Sparse dictionary*: LIA-X [17] showed that L1-regularized sparse activation improves motion interpretability and enables single-factor editing, which we corroborate qualitatively in Fig. 7.

(3) *Divergence loss*: the FDM framework [15] proved that aligning flow divergence reduces the gap between learned and true probability paths; our adoption is motivated by the sensitivity of motion latent distributions to trajectory accuracy.

(4) *Position-shift offset*: Sonic [1] introduced the shifting strategy to expand temporal receptive fields; we set $o{=}7$ to ensure that $N_{\text{fe}}{\times}o{=}70 > T{=}25$, guaranteeing full sequence coverage across ODE steps.

(5) *Long-sequence identity stability*: the CSIM metric explicitly measures identity consistency over time; our method achieves the highest CSIM across both benchmarks, and the long-form qualitative results in Fig. 6 show no visible identity drift over extended sequences.

TABLE II
COMPARISON OF INFERENCE SPEED AND MODEL SIZE. OUR METHOD ACHIEVES THE HIGHEST REAL-TIME EFFICIENCY WITH A COMPACT MODEL SIZE, EVALUATED ON AN NVIDIA A100 GPU.

Method	**FPS** ↑	**Model Size (MB)** ↓
SadTalker [28]	5.75	4097.55
EchoMimic [2]	0.62	4349.78
Hallo [45]	0.47	9499.45
Hallo2 [3]	0.47	9499.45
Ours	**143.82**	**2355.68**

VI. CONCLUSIONS

This paper introduces a voice-driven portrait animation framework built upon flow-divergence matching within an orthogonal latent motion space. By decomposing this space into identity-preservable and motion-transferable subspaces, the model achieves clear disentanglement and produces realistic, consistent facial motion across identities. A sparse motion dictionary further enhances interpretability and enables fine-grained semantic control over facial dynamics. To model the relationship between speech and motion, a transformer-based vector field predictor establishes a coherent mapping from audio cues to temporal evolution, while the flow-based temporal fusion mechanism captures both short-term phoneme dynamics and long-term temporal structure, minimizing motion jitter and discontinuities. Extensive experiments on HDTF and CelebV-HQ validate that our method surpasses previous approaches in quality, expressiveness, and temporal stability, providing a unified and interpretable foundation for generative motion representation learning.

REFERENCES

[1] X. Ji, X. Hu, Z. Xu, J. Zhu, C. Lin, Q. He, J. Zhang, D. Luo, Y. Chen, Q. Lin, Q. Lu, and C. Wang, "Sonic: Shifting focus to global audio perception in portrait animation," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 193–203.

[2] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, "EchoMimic: lifelike audio-driven portrait animations through editable landmark conditions," in *Proc. of AAAI Conference on Artificial Intelligence*, 2025.

[3] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang, "Hallo2: Long-duration and high-resolution audio-driven portrait image animation," in *Proc. of International Conference on Learning Representations*, 2025.

[4] J. Chien and H. Liu, "Contrastive disentanglement learning for empathetic generation," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2025, pp. 1–6.

[5] J.-T. Chien and Z.-X. Tai, "Reinforced retrieval-augmented generation in large language models," in *Proc. of International Joint Conference on Neural Networks*, 2025, pp. 1–7.

[6] H.-C. Yu and J.-T. Chien, "Attention disentanglement for semantic diffusion modeling in text-to-image generation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[7] J. Jiang, G. Lin, Z. Rong, C. Liang, Y. Zhu, J. Yang, and T. Zhong, "MobilePortrait: Real-time one-shot neural head avatars on mobile devices," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 15 920–15 929.

[8] B. Kerbl, G. Kopanas, T. Leimkuehler, and G. Drettakis, "3D gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, pp. 1–14, 2023.

[9] K. Cho, J. Lee, H. Yoon, Y. Hong, J. Ko, S. Ahn, and S. Kim, "Gaussiantalker: Real-time talking head synthesis with 3D Gaussian splatting," in *Proc. of ACM International Conference on Multimedia*, 2024, pp. 10 985–10 994.

[10] J. Li, J. Zhang, X. Bai, J. Zheng, J. Zhou, and L. Gu, "InsTaG: Learning personalized 3d talking head from few-second video," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 10 690–10 700.

[11] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "LIA: Latent image animator," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 829–10 844, 2024.

[12] J. Ling, H. Xue, A. Tang, R. Xie, and L. Song, "ViCoFace: Learning disentangled latent motion representations for visual-consistent face reenactment," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 12, 2024.

[13] Y.-H. Park, M. Kwon, J. Choi, J. Jo, and Y. Uh, "Understanding the latent space of diffusion models through the lens of riemannian geometry," *Advances in Neural Information Processing Systems*, vol. 36, pp. 24 129–24 142, 2023.

[14] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *Proc. of International Conference on Learning Representations*, 2023.

[15] Y. Huang, T. Transue, S.-H. Wang, W. M. Feldman, H. Zhang, and B. Wang, "Improving flow matching by aligning flow divergence," in *Proc. of International Conference on Machine Learning*, 2025.

[16] T. Ki, D. Min, and G. Chae, "Float: Generative motion latent flow matching for audio-driven talking portrait," in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2025, pp. 14 699–14 710.

[17] Y. Wang, D. Yang, X. Chen, F. Bremond, Y. Qiao, and A. Dantcheva, "LIA-X: Interpretable latent portrait animator," *arXiv preprint arXiv:2508.09959*, 2025.

[18] W. S. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2022, pp. 4172–4182.

[19] J.-T. Chien and C.-C. Li, "Learnable contrastive decoding for dehallucination in LLMs," in *Proc. of International Joint Conference on Neural Networks*, 2026.

[20] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9240–9249.

[21] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the GAN latent space," in *Proc. of International Conference on Machine Learning*, 2020.

[22] Z. Li, M.-W. Mak, J.-T. Chien, M. Pilanci, Z. Jin, and H. Meng, "Disentangling speech representations learning with latent diffusion for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3896–3907, 2025.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Neural Information Processing Systems*, 2020, pp. 6840–6851.

[24] T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations." in *Proc. of International Conference on Neural Information Processing Systems*, 2018, pp. 6572–6583.

[25] S. Ploumpis, H. Wang, N. E. Pears, W. Smith, and S. Zafeiriou, "Combining 3D Morphable Models: A large scale face-and-head model," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 926–10 935.

[26] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, 2019.

[27] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[28] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8652–8661.

[29] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo, "VASA-1: lifelike audio-driven talking faces generated in real time," *Advances in Neural Information Processing Systems*, 2025.

[30] C.-C. Chen, H.-Y. Lin, and J.-T. Chien, "CGDD: Contrastive Gaussian-Dirac diffusion model," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[31] T.-C. Luo and J.-T. Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[32] J.-T. Chien and S.-J. Huang, "Learning flow-based disentanglement," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 2, pp. 2390–2402, 2024.

[33] J.-T. Chien and C.-K. Yeh, "Dynamic diffusion programming and classification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 1594–1607, 2026.

[34] J.-T. Chien and Y.-H. Huang, "Latent semantic and disentangled attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 047–10 059, 2024.

[35] H. Lio, S.-E. Li, and J.-T. Chien, "Adversarial mask transformer for sequential learning," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 4178–4182.

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of International Conference on Machine Learning*, 2023.

[37] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[38] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 9168–9178.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[40] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2017.

[41] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2016, pp. 2414–2423.

[42] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4685–4694.

[43] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3660–3669.

[44] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "CelebV-HQ: A large-scale video facial attributes dataset," in *Proc. of European Conference on Computer Vision*, 2022, p. 650–667.

[45] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu, "Hallo: Hierarchical audio-driven visual synthesis for portrait image animation," *ArXiv*, vol. abs/2406.08801, 2024.

[46] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. of IEEE/CVF International Conference on Computer Vision*, 2017, pp. 1021–1030.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[48] Q. Dao, H. Phung, B. Nguyen, and A. Tran, "Flow matching in latent space," *arXiv preprint arXiv:2307.08698*, 2023.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Neural Information Processing Systems*, 2017.

[50] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "FVD: A new metric for video generation," in *Proc. of Diffusion, Generative and Score-based Models Workshop at ICLR*, 2019.

[51] S. E. Ghazouali, U. Michelucci, Y. E. Hillali, and H. Nouira, "CSIM: A copula-based similarity index sensitive to local changes for image quality assessment," *arXiv preprint arXiv:2410.01411*, 2024.

[52] L. Tian, Q. Wang, B. Zhang, and L. Bo, "Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions," in *Proc. of European Conference on Computer Vision*, 2024, pp. 244–260.