

KSDIFF: KEYFRAME-AUGMENTED SPEECH-AWARE DUAL-PATH DIFFUSION FOR FACIAL ANIMATION

Tianle Lyu[†], Junchuan Zhao[†], Ye Wang^{*}

School of Computing, National University of Singapore

ABSTRACT

Audio-driven facial animation has made significant progress in multimedia applications, with diffusion models showing strong potential for talking-face synthesis. However, most existing works treat speech features as a monolithic representation and fail to capture their fine-grained roles in driving different facial motions, while also overlooking the importance of modeling keyframes with intense dynamics. To address these limitations, we propose KSDiff, a Keyframe-Augmented Speech-Aware Dual-Path Diffusion framework. Specifically, the raw audio and transcript are processed by a Dual-Path Speech Encoder (DPSE) to disentangle expression-related and head-pose-related features, while an autoregressive Keyframe Establishment Learning (KEL) module predicts the most salient motion frames. These components are integrated into a Dual-path Motion generator to synthesize coherent and realistic facial motions. Extensive experiments on HDTF and VoxCeleb demonstrate that KSDiff achieves state-of-the-art performance, with improvements in both lip synchronization accuracy and head-pose naturalness. Our results highlight the effectiveness of combining speech disentanglement with keyframe-aware diffusion for talking-head generation. The project page is available at: <https://kincin.github.io/KSDiff/>.

Index Terms— Talking head synthesis, Diffusion models, Keyframe modeling, Head pose and expression dynamics

1. INTRODUCTION

Audio-driven facial animation has attracted increasing attention multimedia due to its wide applications in digital entertainment, virtual avatars, and human-computer interaction. Recently, diffusion models have demonstrated remarkable capability in synthesizing realistic and temporally coherent talking faces. Despite these advances, most existing methods [1, 2, 3, 4] treat speech features as a monolithic representation and overlook their fine-grained roles in driving different facial motions.

Several recent studies have attempted to address these challenges. SadTalker [5] and Synctalk [6] focuses on head-pose modeling but lacks refined speech guidance. ProsodyTalker [7] introduces prosody into generation to explore the role of speech information, but remains unstable under extreme conditions. KeyFace [8] emphasizes keyframe selection, yet the chosen frames often lack clear physical meaning. These observations highlight the necessity of disentangling speech representations and designing principled keyframe modeling strategies.

With the above motivations, we further observe an interesting phenomenon, as illustrated in Fig. 1: speech features related

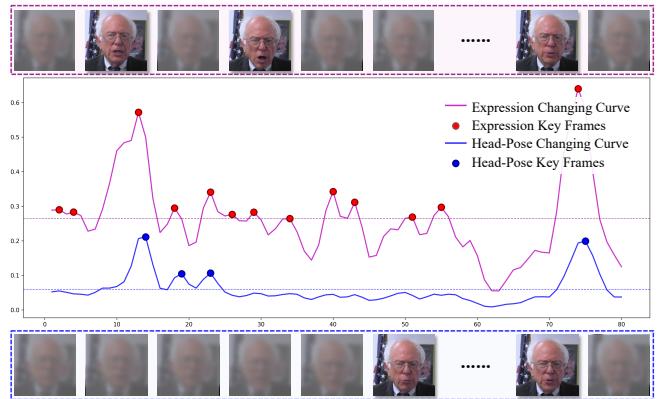


Fig. 1. Motivation illustration. Expression features tend to be more dynamic than head-pose features. Specifically, expressions are associated with high-frequency variations, while head poses primarily reflect low-frequency information.

to expressions typically correspond to high-frequency variations, while head-pose-related features mainly capture low-frequency components. [9] Inspired by the concept of anchor points in KeyFace [8] and the motion-appearance disentanglement strategy in FD2Talk [10], we extend this idea by explicitly leveraging keyframe information during generation. To this end, we propose a novel pipeline, KSDiff, a **Keyframe-Augmented Speech-Aware Dual-Path Diffusion** model for facial animation. Specifically, we design a **Dual-Path Speech Encoder (DPSE)** to disentangle the input audio waveform into expression-related and head-pose-related speech features. These are then processed by the **Keyframe Establishment Learning (KEL)** module to generate two corresponding keyframe sequences. Finally, all extracted conditions are integrated into a **DiffSpeaker** [3] based Dual-Path Motion Generator to synthesize coherent facial motion that jointly captures both expressions and head-pose. Extensive experiments demonstrate that KSDiff achieves state-of-the-art performance across multiple benchmarks.

The main contributions are summarized as follows:

- We propose a Dual-Path Speech Encoder (DPSE) that disentangles speech features into expression-related and head-pose-related components, facilitating the synthesis of each motion type with precise feature representations.
- We introduce a Keyframe Establishment Learning (KEL) that ensures frames with the most intense movements are selected, thereby improving the fidelity of talking.
- The proposed dual-path diffusion framework produces highly detailed facial animations with realistic motion dynamics, and extensive experiments validate its effectiveness.

[†] Equal contribution.

^{*} Corresponding author. Email: wangye@comp.nus.edu.sg.
Emails: tianle_lyu@u.nus.edu, junchuan@u.nus.edu.

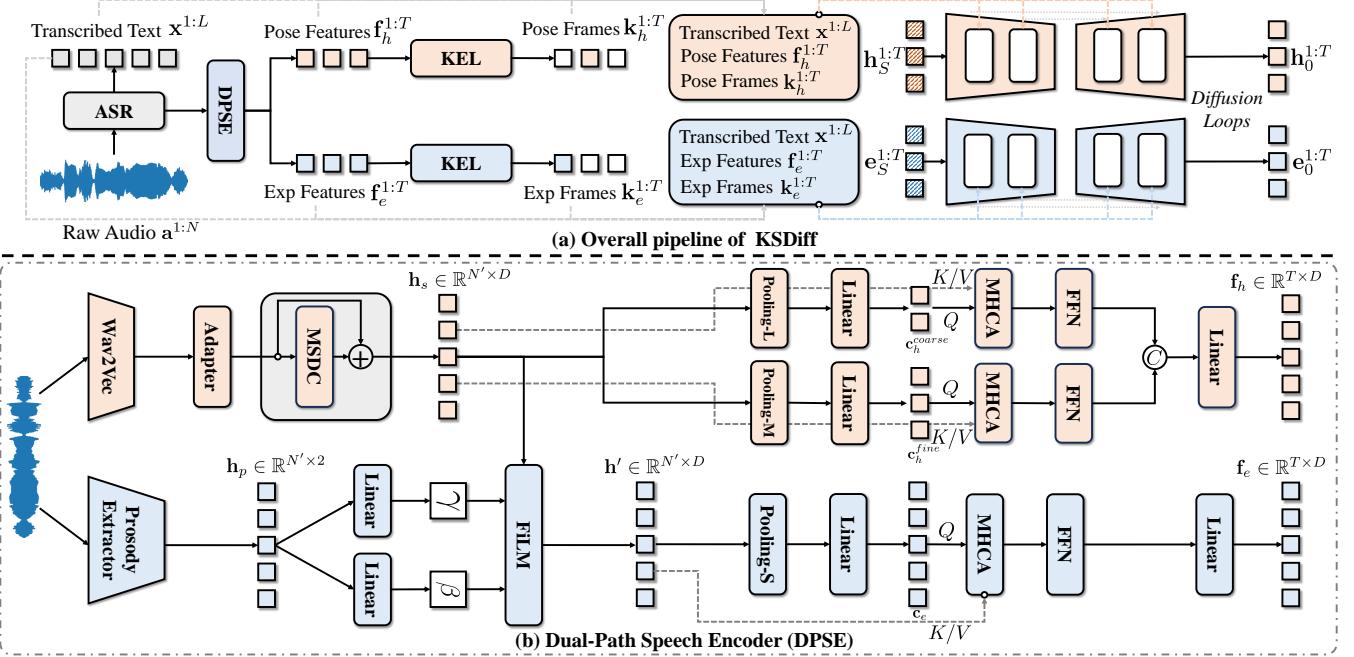


Fig. 2. Overview of KSDiff. Input audio is processed by the **Dual-Path Speech Encoder** (DPSE) to disentangle expression- and head-pose-related features. The **Keyframe Establishment Learning** (KEL) module extracts corresponding keyframe sequences, which together with the disentangled features are fed into Dual-Path Motion Generators to produce head-pose and expression coefficients.

2. METHODOLOGY

The overall framework is illustrated in Fig. 2. Given raw audio $a^{1:N}$ and transcribed text $x^{1:L}$, we first employ a Dual-Path Speech Encoder (DPSE) to disentangle head-pose-related features $f_h^{1:T}$ and expression-related features $f_e^{1:T}$. Together with the transcript $x^{1:L}$, these features are processed by the Keyframe Establishment Learning (KEL) module, which generates keyframe sequences $k_h^{1:T}$ and $k_e^{1:T}$. Subsequently, all components are passed to the DiffSpeaker-based Dual-Path Motion Generator [3], which predicts head-pose coefficients $h^{1:T}$ and expression coefficients $e^{1:T}$. Finally, DECA [11] renders the complete talking-head motion $m^{1:T}$.

2.1. Dual-Path Speech Encoder (DPSE)

Motivated by the observation that expression-related speech cues correspond to high-frequency variations, whereas head-pose-related cues are dominated by low-frequency components [12], we propose a Dual-Path Speech Encoder (DPSE) that explicitly separates speech into two parts [13]. Given a raw waveform $a^{1:N}$, a frozen Wav2Vec encoder $f_{SE}(\cdot)$ [14] extracts frame-level features, which are adapted by a lightweight projection layer. To capture temporal structures at multiple scales with minimal overhead, we employ a parallel Multi-Scale Dilated Convolution (MSDC) block. The MSDC consists of L branches with dilations $\{d_\ell\}_{\ell=1}^L$, where each branch contains a depthwise k -convolution, GroupNorm, GLU gating, and a pointwise convolution. Branch outputs are concatenated, projected by a convolution layer, and added residually to produce $h_s \in \mathbb{R}^{N' \times D}$.

The hidden speech features h_s are split into two branches, one targeting expression-related cues and the other focusing on head-pose information. For the head-pose branch, we apply two windowed pooling operations with different receptive fields: a long

window w_h^c and a mid-sized window w_h^f , yielding multi-resolution speech representations. Each sequence is further mapped by linear projections to produce coarse and fine head-pose features, denoted as $c_h^{coarse} \in \mathbb{R}^{N_h^c \times D}$ and $c_h^{fine} \in \mathbb{R}^{N_h^f \times D}$.

In parallel, motivated by findings that prosody strongly correlates with expression features [7, 15, 16], we directly extract f_0 and energy from the raw waveform $a^{1:N}$, with per-utterance normalization for robustness. The resulting prosody representation $h_p \in \mathbb{R}^{N' \times 2}$ is passed through two parallel linear layers to generate FiLM [17] conditioning parameters: a scaling factor γ and a bias term β . These parameters modulate h_s through a FiLM operation to obtain the prosody-aware speech features $h' \in \mathbb{R}^{N' \times D}$, after which the modulated features are pooled with a short window w_e and passed through a linear layer, yielding the expression-related token sequence $c_e \in \mathbb{R}^{N_e \times D}$.

Subsequently, the speech features with different receptive fields are refined through multi-head cross-attention (MHCA), enabling dynamic interaction across scales. Let $\mathbf{C} = \{c_h^{coarse}, c_h^{fine}, c_e\}$ denote the multi-scale speech features. For each branch $\mathbf{X}_i \in \{\mathbf{X}_h^{coarse}, \mathbf{X}_h^{fine}, \mathbf{X}_e\}$, the corresponding queries, keys, and values are defined as:

$$\mathbf{Q}_i = \mathbf{W}_i^Q \mathbf{C}_i, \quad \mathbf{K}_i = \mathbf{W}_i^K \mathbf{H}_i, \quad \mathbf{V}_i = \mathbf{W}_i^V \mathbf{H}_i, \quad (1)$$

where $\mathbf{H} = \{h_s, h', h\}$ are the hidden speech features aligned with the respective branches, and \mathbf{W}_i^Q , \mathbf{W}_i^K , and \mathbf{W}_i^V are the learnable parameter matrices. The cross-attention output for each branch is computed as:

$$\mathbf{o}_i = \text{softmax} \left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d}} \right) \mathbf{V}_i. \quad (2)$$

The outputs of MHCA are subsequently passed through feed-forward networks (FFN) to further enhance representational capac-

ity. The coarse and fine head-pose features are then concatenated and projected via a linear layer to form the head-pose-related representation $\mathbf{f}_h \in \mathbb{R}^{T \times D}$, while the expression branch produces the expression-related representation $\mathbf{f}_e \in \mathbb{R}^{T \times D}$.

2.2. Keyframe Establishment Learning (KEL)

Inspired by KeyFace [8], we highlight the importance of keyframes in talking-head generation. To identify key motion moments, we measure the inter-frame variation of ground-truth head-pose $\hat{\mathbf{h}} \in \mathbb{R}^{T \times 9}$ and expression parameters $\hat{\mathbf{e}} \in \mathbb{R}^{T \times 50}$.

The head-pose $\hat{\mathbf{h}}$ is decomposed into pose parameters \mathbf{p} and camera parameters \mathbf{c} , such that $\hat{\mathbf{h}} = [\mathbf{p}, \mathbf{c}]$, $\mathbf{p} \in \mathbb{R}^{T \times 6}$, $\mathbf{c} \in \mathbb{R}^{T \times 3}$. Within \mathbf{p} , the first three dimensions correspond to the rotation parameters $\mathbf{r} \in \mathbb{R}^{T \times 3}$. The relative rotation between consecutive frames is defined as $\Delta r_t = \mathbf{r}_t \mathbf{r}_{t-1}^\top$, from which the angular magnitude θ_t is extracted. The remaining three dimensions of \mathbf{p} represent neck parameters $\mathbf{n} \in \mathbb{R}^{T \times 3}$, which are concatenated with the camera parameters \mathbf{c} to form the combined sequence $\mathbf{c}' = [\mathbf{n}, \mathbf{c}] \in \mathbb{R}^{T \times 6}$. The inter-frame variation of these parameters is measured using the Euclidean distance $\Delta c'_t = \|\mathbf{c}'_t - \mathbf{c}'_{t-1}\|_2$.

For expression coefficients $\hat{\mathbf{e}}$, the variation is directly computed as the inter-frame Euclidean distance. The final overall head-pose and expression variation sequences are then given by:

$$\delta_{h,t} = \theta_t + \Delta c'_t, \quad \delta_{e,t} = \|\hat{\mathbf{e}}_t - \hat{\mathbf{e}}_{t-1}\|_2. \quad (3)$$

Both sequences $\delta^h \in \mathbb{R}^T$ and $\delta^e \in \mathbb{R}^T$ are smoothed using a Gaussian filter, and local maxima above a data-dependent threshold are selected as target head-pose keyframes $\hat{\mathbf{k}}_h \in \{0, 1\}^T$ and expression keyframes $\hat{\mathbf{k}}_e \in \{0, 1\}^T$.

To predict key moments, we employ two Transformer-based predictors that autoregressively generate binary keyframe sequences conditioned on speech embeddings and text features. Given the processed speech features \mathbf{f}_h , \mathbf{f}_e and the transcribed text \mathbf{x} , the predictors produce head-pose keyframes \mathbf{k}_h and expression keyframes \mathbf{k}_e . To mitigate the strong class imbalance between sparse keyframes and abundant non-keyframes, we adopt a weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}}^h = - \sum_t \left(w_1 \hat{k}_{h,t} \log k_{h,t} + w_0 (1 - \hat{k}_{h,t}) \log (1 - k_{h,t}) \right), \quad (4)$$

where $k_{h,t} \in \{0, 1\}$ is the predicted probability that frame t is a head-pose keyframe and $\hat{k}_{h,t} \in \{0, 1\}$ is the ground truth. The loss for expression keyframes, $\mathcal{L}_{\text{BCE}}^e$, follows the same formulation. The positive class (keyframe) is assigned a larger weight $w_1 > w_0$, while the negative class (non-keyframe) is assigned w_0 .

2.3. Dual-Path Motion Generator

Based on DiffSpeaker [3], we design a dual-path framework to separately generate head-pose and expression coefficients, denoted as $\mathbf{h}^{1:T}$ and $\mathbf{e}^{1:T}$. In the head-pose path, the diffusion process is conditioned on the transcribed text $\mathbf{x}^{1:L}$, head-pose-related speech features $\mathbf{f}_h^{1:T}$, and the head-pose keyframe sequence $\mathbf{k}_h^{1:T}$. In parallel, the expression path employs the same conditioning scheme with expression-related speech features $\mathbf{f}_e^{1:T}$ and keyframe sequence $\mathbf{k}_e^{1:T}$. Following the DiffSpeaker formulation, the diffusion loss for each path is defined as:

$$\mathcal{L}_{\text{diff}}^h = \lambda_1 \mathcal{L}_{\text{rec}}^h + \lambda_2 \mathcal{L}_{\text{vel}}^h, \quad (5)$$

where $\lambda_1 = \lambda_2 = 1$, and $\mathcal{L}_{\text{diff}}^e$ follows an identical formulation.

To further enhance motion quality, we incorporate a multi-resolution spectral loss and a dynamics regularization term. Specifically, we adapt the MR-STFT loss [18] to kinematic sequences and apply it independently to each branch:

$$\mathcal{L}_{\text{mr}}^h = \sum_{r \in \mathcal{R}} \| |S_r(\mathbf{h})| - |S_r(\hat{\mathbf{h}})| \|_1, \quad (6)$$

with $\mathcal{L}_{\text{mr}}^e$ defined analogously.

Finally, the total loss of the proposed KSDiff framework is expressed as:

$$\begin{aligned} \mathcal{L}_h &= \lambda_{\text{mr}}^h \mathcal{L}_{\text{mr}}^h + \lambda_{\text{BCE}}^h \mathcal{L}_{\text{BCE}}^h + \lambda_{\text{diff}}^h \mathcal{L}_{\text{diff}}^h, \\ \mathcal{L}_e &= \lambda_{\text{mr}}^e \mathcal{L}_{\text{mr}}^e + \lambda_{\text{BCE}}^e \mathcal{L}_{\text{BCE}}^e + \lambda_{\text{diff}}^e \mathcal{L}_{\text{diff}}^e, \end{aligned} \quad (7)$$

where \mathcal{L}_h and \mathcal{L}_e denote the losses for the head-pose and expression branches, respectively. We set the loss weights as $\lambda_{\text{mr}} = 0.3$, $\lambda_{\text{BCE}} = 0.5$, and $\lambda_{\text{diff}} = 1$ for both branches.

3. EXPERIMENTS SETUPS

3.1. Dataset

We train and evaluate our model on two benchmarks. The High-Definition Talking Face (HDTF) dataset [19] contains high-quality frontal talking-face clips with diverse expressions, making it a standard benchmark. In contrast, the VoxCeleb dataset [20] includes large-scale speaker videos collected in unconstrained conditions with significant variations in pose, background, and recording quality, serving to evaluate the generalization of our approach.

3.2. Evaluation Metrics

We evaluate our model with metrics covering both lip synchronization and head-pose motion. For lip synchronization, *LVE* (Lip Vertex Error) [22] measures the Euclidean distance between predicted and ground-truth lip vertices. *LSE-D* and *LSE-C* [23] come from a pre-trained lip-sync discriminator: *LSE-D* quantifies audio-video embedding distance, while *LSE-C* reflects discriminator confidence. For head pose, *Diversity* [24] captures the variance of head-pose trajectories, and *Beat Align* [25] measures alignment between motion beats and speech accents. Together, these metrics comprehensively assess lip accuracy and head dynamics.

3.3. Implementation Details

For both datasets, we preprocess videos with DECA [11] to obtain per-frame expression and head-pose coefficients $\hat{\mathbf{e}} \in \mathbb{R}^{T \times 50}$ and $\hat{\mathbf{h}} \in \mathbb{R}^{T \times 9}$. In parallel, `ffmpeg` is used to extract raw audio, Whisper [26] provides transcribed text \mathbf{x} , and prosody features \mathbf{h}_p are obtained as described in Sec. 2.1. Keyframe sequences are extracted according to the method in Sec. 2.2. All videos are aligned to faces, and audio is resampled to 16kHz.

In the DPSE module, we set the hidden dimension $D = 512$, kernel size $k = 5$, and apply MSDC followed by a dropout rate of 0.1. The stride s_n is chosen from $\{2, 4\}$, and the fused feature dimension is $d_c = 512$. We use $w_h^c = 1.0$, $w_h^f = 0.25$, and $w_e = 0.1$. We use the AdamW optimizer for 100k iterations, with 5k warmup steps, a batch size of 32, and a learning rate of $1e-4$. The hidden feature dimension is set to 512, and the transformer decoder in the keyframe branch has 6 layers and 8 attention heads. The overall training on four NVIDIA RTX A5000 GPUs takes about 16 hours.

Table 1. Objective comparison of lip synchronization and head motion on the HDTF [19] and VoxCeleb [20] datasets. Best scores are shown in **bold** and the second best are underlined. For clarity, LVE values are scaled by 10^{-5} mm.

Method	HDTF dataset					VoxCeleb dataset				
	LSE-C↑	LSE-D↓	LVE↓	Diversity↑	Beat Align↑	LSE-C↑	LSE-D↓	LVE↓	Diversity↑	Beat Align↑
SadTalker [5]	0.625	10.121	5.918	0.246	0.274	0.653	9.981	5.802	0.296	0.305
FaceDiffuser [1]	0.594	11.156	6.226	-	-	0.627	10.530	6.091	-	-
DiffTalk [2]	0.689	9.884	5.279	0.281	0.295	0.706	9.743	5.026	0.297	0.324
HalIo2 [21]	0.704	9.629	5.437	<u>0.293</u>	0.302	0.711	9.841	5.174	<u>0.316</u>	0.347
KeyFace [8]	0.717	<u>9.541</u>	5.095	0.274	<u>0.331</u>	0.732	<u>9.415</u>	4.821	0.310	<u>0.354</u>
DiffSpeaker [3]	0.702	9.916	4.926	-	-	0.707	9.732	<u>4.684</u>	-	-
KSDiff (Ours)	0.708	9.204	4.835	0.318	0.354	<u>0.713</u>	9.037	4.327	0.328	0.377

Table 2. Subjective evaluation results on full-face quality, lip synchronization, head motion, and fluency.

Methods	Full-face↑	Lip sync↑	Head motion↑	Fluency↑
SadTalker [5]	3.77	3.64	4.06	3.62
FaceDiffuser [1]	3.24	3.36	1.58	3.37
DiffTalk [2]	4.06	3.91	4.27	4.16
HalIo2 [21]	4.05	4.31	3.98	3.94
KeyFace [8]	4.12	4.24	4.42	4.27
DiffSpeaker [3]	3.69	4.32	1.37	3.63
KSDiff (Ours)	4.22	4.48	4.60	4.45

Table 3. Architecture ablation study on the HDTF dataset [19]. For clarity, LVE values are scaled by 10^{-5} mm.

Methods	LSE-C↑	LSE-D↓	LVE↓	Diversity↑	Beat Align↑
w/o speech split	0.640	9.865	5.445	0.238	0.261
w/o dual-path diff	0.652	9.629	5.172	0.270	0.316
w/o keyframe	0.663	9.570	5.329	0.256	0.292
w/o prosody	0.683	9.481	4.818	0.296	0.331
w/o transcript	0.699	9.372	5.720	0.305	0.342
wav2vec only	0.576	10.528	5.584	0.221	0.254
Ours	0.708	9.204	4.635	0.318	0.354

4. EXPERIMENTS RESULTS

As shown in Table 1, we compare our KSDiff with other state-of-the-art methods in two categories on HDTF dataset [19] and VoxCeleb dataset [20]. We use DiffSpeaker [3] as our baseline model, which adopts a diffusion-based Transformer architecture with biased conditional self- and cross- attention mechanisms for speech-driven 3D facial animation. The results highlight the strong capability of KSDiff in capturing fine-grained expression and head-pose details.

To assess perceptual quality, we conduct a user study on four aspects: 1) *Full-face naturalness*, 2) *Lip-sync accuracy*, 3) *Head motion plausibility*, and 4) *Overall fluency*. We randomly sample generated videos from different methods and ask 26 participants to rate each aspect on a 5-point Likert scale (1 = very poor, 5 = excellent). As shown in Table 2, our KSDiff achieves the highest average scores across all criteria, confirming its superior perceptual quality.

We conduct ablation studies under six settings: 1) *w/o speech split*: use the entire speech feature for both branches without disentangling expression-related and head-pose-related components; 2) *w/o dual-path diffusion*: use a single diffusion process for all conditions, directly fusing expression and head-pose information; 3) *w/o keyframe*: omit the keyframe extraction module; 4) *w/o prosody*: without prosody guidance in the DPSE module; 5) *w/o transcript*: without transcript guidance in the pipeline; 6) *Wav2Vec only*: use raw Wav2Vec features without any refinement modules. As shown

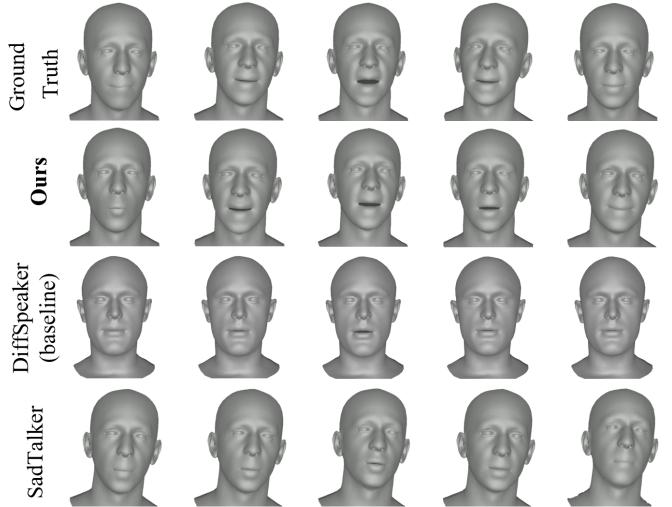


Fig. 3. Visualization comparison with DiffSpeaker (baseline) [3] and SadTalker [5]. All speakers are uttering the word “bread”. Compared to prior methods, our KSDiff generates more natural expressions and richer head-pose dynamics, closely matching the ground-truth sequence.

in Table 3, the results indicate that each component contributes significantly to the overall performance.

As shown in Fig. 3, KSDiff achieves more accurate head-motion trajectories and natural expressions compared with prior methods. In particular, at key phoneme frames, our model generates plausible head rotations rather than the exaggerated dynamics observed in SadTalker [5]. This leads to more faithful expression–pose coordination and overall results that better align with the ground truth.

5. CONCLUSION

In this paper, we propose KSDiff, a keyframe-augmented speech-aware dual-path diffusion framework for audio-driven facial animation. By disentangling speech into expression- and pose-related features and introducing an autoregressive keyframe learning module, our approach produces natural and coherent facial motions. Experiments on HDTF and VoxCeleb demonstrate that KSDiff achieves state-of-the-art performance in both objective metrics and perceptual quality, while ablation studies validate the contribution of each component. These results highlight the effectiveness and versatility of KSDiff in advancing audio-driven facial animation.

6. REFERENCES

[1] S. Stan, K. I. Haque, and Z. Yumak, “Facediffuser: Speech-driven 3d facial animation synthesis using diffusion,” in *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2023, pp. 1–11.

[2] S. Shen, W. Zhao, Z. Meng, W. Li, Z. Zhu, J. Zhou, and J. Lu, “DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1982–1991.

[3] Z. Ma, X. Zhu, G. Qi, C. Qian, Z. Zhang, and Z. Lei, “DiffSpeaker: Speech-driven 3d facial animation with diffusion transformer,” *arXiv preprint arXiv:2402.05712*, 2024.

[4] H. Wang, Y. Weng, Y. Li, Z. Guo, J. Du, S. Niu, J. Ma, S. He, X. Wu, Q. Hu, B. Yin, C. Liu, and Q. Liu, “Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11–15, 2025*, 2025, pp. 26212–26221.

[5] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, and Y. Shan, “Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8652–8661.

[6] Z. Peng, W. Hu, Y. Shi, X. Zhu, X. Zhang, H. Zhao, J. He, H. Liu, and Z. Fan, “SyncTalk: The devil is in the synchronization for talking head synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 666–676.

[7] Z. Li, X. Lv, Q. Liu, Q. Meng, X. Sun, and S. Zhang, “Prosodytalker: 3d visual speech animation via prosody decomposition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, vol. 39, pp. 5110–5118.

[8] A. Bigata, M. Stypulkowski, R. Mira, S. Bounareli, K. Vougioukas, Z. Landgraf, N. Drobyshev, M. Zieba, S. Petridis, and M. Pantic, “Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5477–5488.

[9] C. Cai, G. Guo, J. Li, J. Su, F. Shen, C. He, J. Xiao, Y. Chen, L. Dai, and F. Zhu, “Speak: Speech-driven pose and emotion-adjustable talking head generation,” in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[10] Z. Yao, X. Cheng, and Z. Huang, “Fd2Talk: Towards generalized talking head generation with facial decoupled diffusion model,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3411–3420.

[11] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, “Learning an animatable detailed 3d face model from in-the-wild images,” *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[12] G. Hwang, S. Hong, S. Lee, S. Park, and G. Chae, “DiscoHead: audio-and-video-driven talking head generation by disentangled control of head pose and facial expressions,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] X. Liu, Z. Liu, and C. Bi, “Nerf-3dtalker: Neural radiance field with 3d prior aided audio disentanglement for talking head synthesis,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[14] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised pre-training for speech recognition,” in *Interspeech 2019*, 2019, pp. 3465–3469.

[15] J. Zhao, C. Low, and Y. Wang, “Spsinger: Multi-singer singing voice synthesis with short reference prompt,” in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[16] J. Zhao, X. Wang, and Y. Wang, “Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning,” in *Interspeech 2025*, 2025, pp. 4893–4897.

[17] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[18] E. Song, R. Yamamoto, M.-J. Hwang, J.-S. Kim, O. Kwon, and J.-M. Kim, “Improved parallel wavegan vocoder with perceptually weighted spectrogram loss,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 470–476.

[19] 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.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: A large-scale speaker identification dataset,” in *Interspeech 2017*, 2017, pp. 2616–2620.

[21] 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*, 2025.

[22] A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh, “Meshtalk: 3d face animation from speech using cross-modality disentanglement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1173–1182.

[23] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. 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*, 2020, pp. 484–492.

[24] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[25] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu, “Bailando: 3d dance generation by actor-critic gpt with choreographic memory,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11050–11059.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavy, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International conference on machine learning*, 2023, pp. 28492–28518.