

# DualTalk: Dual-Speaker Interaction for 3D Talking Head Conversations

## Supplementary Material

In this supplementary material, we provide additional details on DualTalk. Section 1 covers the implementation details, including network architecture and loss functions. Section 2 describes the dataset collection and processing methods. Section 3 outlines the evaluation metrics used to assess performance. Section 4 discusses ethical considerations, and Section 5 addresses limitations and future work.

### 1. Implementation Details

#### 1.1. Network Architecture

In this section, we provide comprehensive implementation details of our DualTalk framework. The framework consists of four main components: Dual-Speaker Joint Encoder, Cross-Modal Temporal Enhancer, Dual-Speaker Interaction Module, and Expressive Synthesis Module.

The Dual-Speaker Joint Encoder processes both audio and visual inputs through parallel branches. For audio processing, we utilize a pre-trained Wav2Vec 2.0 [1] model to encode the raw audio waveforms (sampled at 16kHz) into high-dimensional feature representations. The audio encoder consists of 12 transformer [10] layers with a hidden dimension of 1024, followed by a linear projection layer that maps the features to a 256-dimensional space. This projection is essential for aligning the audio features with the visual representation space. The visual branch processes blendshape coefficients through a two-layer MLP with ReLU activations, where the first layer maps the 56 blendshape parameters to 128 dimensions, and the second layer further projects these features to match the 256-dimensional audio features.

The Cross-Modal Temporal Enhancer is designed to ensure temporal coherence and modal alignment. At its core is a multimodal cross-attention mechanism with 4 attention heads. This mechanism allows the model to establish connections between audio and visual features across different temporal positions. Following the cross-attention layer, we employ a bidirectional LSTM [5] with 512 hidden units and 2 layers to capture long-term dependencies in both forward and backward directions. The LSTM incorporates a dropout of 0.1 between layers to prevent overfitting.

For the Dual-Speaker Interaction Module, we implement a transformer-based architecture consisting of an encoder and decoder, each with 3 layers. The encoder employs 4-head self-attention mechanisms with a hidden dimension of 256 and a feed-forward network dimension of 512. The Modal Alignment Attention layer, inspired by FaceFormer [3], uses a custom attention mask to ensure causal relationships in the temporal domain. The decoder

follows a similar structure but includes additional cross-attention layers to integrate information from both speakers.

The Expressive Synthesis Module utilizes an adaptive expression modulation mechanism implemented as a two-layer MLP. The first layer expands the 256-dimensional features to 512 dimensions, followed by layer normalization and ReLU activation. The second layer then projects back to 256 dimensions before the final blendshape prediction layer, which outputs 56 blendshape parameters normalized through a sigmoid activation.

#### 1.2. Loss Functions

Our training objective incorporates multiple loss terms to ensure both accurate blendshape prediction and smooth temporal dynamics. The total loss function consists of two primary components: a direct blendshape reconstruction loss and a velocity loss that enforces temporal consistency.

The blendshape reconstruction loss ( $\mathcal{L}_{bs}$ ) is computed as the Mean Squared Error (MSE) between the predicted head motion blendshape parameters ( $\hat{M}$ ) and the ground truth blendshapes ( $M$ ):

$$\mathcal{L}_{bs} = \text{MSE}(\hat{M}, M) = \frac{1}{N} \sum_{i=1}^N (\hat{M}_i - M_i)^2 \quad (1)$$

To ensure smooth and natural facial movements, we introduce a velocity loss term that penalizes sudden changes in blendshape parameters between consecutive frames. The velocity is computed as the first-order temporal difference of blendshape parameters. Specifically, for both predicted and ground truth sequences, we calculate the frame-to-frame differences:

$$V_{gt} = M_{t+1} - M_t \quad (2)$$

$$\hat{V} = \hat{M}_{t+1} - \hat{M}_t \quad (3)$$

where  $t$  represents the frame index. The velocity loss ( $\mathcal{L}_{vel}$ ) is then computed as the MSE between the predicted and ground truth velocities:

$$\mathcal{L}_{vel} = \text{MSE}(\hat{V}, V_{gt}) = \frac{1}{N-1} \sum_{t=1}^{N-1} (\hat{V}_t - V_{gt,t})^2 \quad (4)$$

The final loss is the mean of these two components:

$$\mathcal{L}_{total} = \mathcal{L}_{bs} + \mathcal{L}_{vel} \quad (5)$$

This combined loss function effectively balances between accurate facial expression reproduction and temporal smoothness. The blendshape reconstruction loss ensures that the predicted facial expressions match the ground truth at each frame, while the velocity loss prevents unrealistic, jittery movements by encouraging smooth transitions between consecutive frames. During training, we use equally weighting these two terms (with an implicit weight of 1.0 for each).

### 1.3. Training Details

During training, we optimize our model using the Adam [6] optimizer with an initial learning rate of  $1e-4$ . We train the model for 200 epochs using a batch size of 32 on a NVIDIA A6000 GPU with 48GB memory each. The complete training process takes approximately 48 hours to converge.

## 2. Dataset Details

Our dataset collection and processing pipeline is designed to create a comprehensive and high-quality dataset for dual-speaker interaction modeling. Here, we provide detailed information about our data collection, processing procedures, and dataset statistics.

The raw data is collected from YouTube interviews, with a wide variety of natural face-to-face interactions. We specifically focus on videos featuring clear facial visibility of both speakers, high-quality audio, and natural conversational dynamics. All videos are in 1920×1080 resolution recorded at 25 frames per second, with audio sampled at 16kHz. The collected videos span different languages, speaking styles, and environmental conditions to ensure robustness and generalization of our model.

The resulting dataset comprises 50 hours of processed conversation data, featuring 1,052 unique identities across 5,858 video clips. Each clip contains an average of 2.5 conversation rounds, where speakers naturally alternate between speaking and listening roles. The dataset is carefully split into training (4,935 clips), testing (539 clips), and out-of-distribution (OOD) validation sets (384 clips). The OOD set specifically includes speakers and conversation scenarios not present in the training data to evaluate generalization capability.

To construct this dataset, we sourced two-person conversational videos from YouTube and RealTalk [4] raw videos. Videos are segmented using TransNet V2 [9] for shot transition detection, retaining only segments longer than 5 seconds to capture meaningful interactions. Visual-guided speech separation is performed with IIANet [7], producing isolated audio streams for each speaker—a critical feature for accurate lip synchronization and expression modeling.

To ensure speaker-specific frame isolation, we use MediaPipe [8] for face detection and tracking. High-resolution 3D facial meshes are extracted using Spectre, and samples

with abnormal coefficients are filtered out. For speaker separation, Pyannote [2] is employed, allowing the identification of multi-round conversations and distinct speaker turns to facilitate the extraction of back-and-forth dialogues. To ensure annotation stability, a minimum speech duration of 2 seconds is set.

## 3. Evaluation Metrics

In this section, we provide detailed descriptions of the evaluation metrics used to assess the performance of our DualTalk framework. These metrics are carefully selected to comprehensively evaluate different aspects of the generated conversational animations, including motion realism, temporal synchronization, and interaction dynamics.

**Fréchet Distance (FD):** The FD serves as our primary metric for evaluating motion realism. It computes the distributional distance between generated and ground-truth motions in the feature space. Specifically, we extract deep features from both the predicted and actual motion sequences using a pre-trained motion encoder, modeling them as multivariate Gaussian distributions. The FD effectively captures the statistical similarity between the generated and real motion distributions, where a lower score indicates better motion realism.

**Paired Fréchet Distance (P-FD):** To evaluate the quality of dual-speaker interactions, we introduce the P-FD metric, which extends the traditional FD by considering the joint distribution of dual-speaker pairs. By concatenating the generated Speaker-B’s motions with the corresponding Speaker-A’s motions along the feature dimension, we compute the FD between these paired representations and their ground-truth counterparts. This approach captures the synchronization and coherence between the two speakers’ movements, providing insights into the quality of interactive dynamics.

**Mean Squared Error (MSE):** For direct motion accuracy assessment, we employ the MSE between generated and ground-truth motions. This metric is computed across all blendshape parameters and temporal dimensions, providing a straightforward measure of prediction accuracy. The MSE helps us understand how closely the generated animations match the ground truth at a frame-by-frame level.

**SI for Diversity (SID):** To evaluate the diversity of generated animations, we use the SID metric. This approach applies k-means clustering ( $k=40$ ) to the motion sequences in the feature space and quantifies diversity by calculating the entropy of the cluster assignment histogram. A higher SID value indicates more diverse and varied motion patterns in the generated animations, which is crucial for producing natural and non-repetitive conversational behaviors.

**Residual Pearson Correlation Coefficient (rPCC):** To assess the temporal correlation between Speaker-A and Speaker-B movements, we introduce the rPCC metric.

It computes the frame-wise Pearson correlation between Speaker-A and Speaker-B motions and then measures the L1 distance between the correlation patterns of generated and ground-truth sequences. The rPCC is particularly useful for evaluating how well the model captures the subtle interactive dynamics between Speaker-A and Speaker-B in conversation.

These metrics collectively provide a comprehensive evaluation framework for assessing the quality, realism, and interactive dynamics of our dual-speaker animation system. Each metric focuses on a specific aspect of the generated animations, enabling detailed analysis of the model’s performance across different dimensions. Through this multifaceted evaluation approach, we can thoroughly validate the effectiveness of our proposed method in generating realistic and interactive conversational animations.

## 4. Ethics Considerations

The development of DualTalk raises important ethical considerations, particularly regarding privacy, misuse, and potential societal impacts. The DualTalk dataset includes extensive conversational data, and while publicly available sources were used, ensuring compliance with data privacy laws and ethical guidelines remains a priority. Steps have been taken to anonymize and process data responsibly, but future work will aim to establish more robust safeguards to prevent inadvertent exposure of personal information.

Another key concern is the potential misuse of DualTalk for deceptive purposes, such as creating realistic yet fabricated conversations or impersonating individuals. To mitigate this, strict usage policies and watermarking techniques can be implemented to differentiate generated content from real-world interactions. Open-sourcing the technology will be accompanied by clear guidelines to discourage unethical applications.

## 5. Limitations and Future Works

The limitations of DualTalk primarily lie in its current focus on dyadic interactions and the lack of precise emotional controllability in generated animations. While DualTalk excels in creating synchronized and natural two-speaker conversations, it cannot yet handle multi-party interactions, which are common in real-world applications. Additionally, while the Expressive Synthesis Module generates nuanced facial expressions, the model lacks the ability to precisely control the emotional tone of its outputs, limiting its adaptability to specific scenarios or user preferences.

Future work will focus on extending DualTalk to multi-party interactions, enabling the model to handle dynamic role transitions and conversational flows in group settings. Additionally, efforts will be directed toward generating controllable emotions, allowing the system to adapt its re-

sponses to specific emotional tones or user preferences, further enhancing the naturalness and versatility of 3D talking head animations.

## References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020. 1
- [2] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE, 2020. 2
- [3] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022. 1
- [4] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*, 2023. 2
- [5] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997. 1
- [6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 2
- [7] Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. Iianet: An intra-and inter-modality attention network for audio-visual speech separation. In *Forty-first International Conference on Machine Learning*, 2024. 2
- [8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2
- [9] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 2
- [10] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1