

Supplementary Material

Anonymous submission

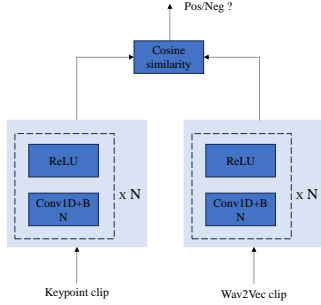


Figure 1: The structure of Adaptive Lip-motion Synchronization Expert

Method

Details of Adaptive Lip-motion Synchronization Expert

Architecture Our synchronization expert takes as input a sequence of T_l consecutive lip keypoint frames and an audio feature clip of size $T_a \times D$, where T_l and T_a represent the number of frames for keypoints and audio, respectively, and D denotes the dimensionality of Wav2Vec features (Schneider et al. 2019). The expert aims to determine whether the input motion and audio are temporally aligned. As illustrated in Fig. 1, it consists of two parallel encoders for landmarks and audio, each built from a stack of 1D convolutions followed by batch normalization and ReLU activation. The training objective uses a cosine-similarity-based binary cross-entropy loss. Specifically, cosine similarity is computed between the keypoints embedding l and the audio embedding a to supervise synchronization.

$$\mathcal{L}_{sync} = CE \left(\frac{a \cdot l}{\max(\|a\|_2 \cdot \|l\|_2, \epsilon)} \right) \quad (1)$$

Training details We train the Adaptive Lip-motion Synchronization Expert (ALSE) on a single NVIDIA H20 GPU, with a batch size of 512. The model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} , which is decayed by a factor of 0.02. Training is performed for a total of 30,000 steps. The training dataset is kept

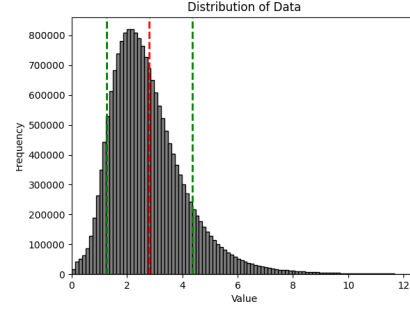


Figure 2: Distribution of motion intensities in the training dataset.

consistent with that used for MoDA, with data distribution across different motion intensities illustrated in Fig. 2.

Method	FVD \uparrow	Sync-C \uparrow	Sync-D \downarrow	RTF \downarrow
Our-s10	205.442	5.804	8.182	0.846
Our-s20	219.762	5.855	8.205	1.048
Our-s30	213.908	5.751	8.295	1.259
Our-s40	213.483	5.741	8.297	1.430
Our-s50	214.182	5.757	8.346	1.635

Table 1: Comparison of MoDA with different inference steps on the CelebV-HQ test set.

Real-time Inference

All experiments were run on one NVIDIA H20 GPU, whose inference speed is comparable to that of an NVIDIA RTX 3090. Relative to VAE-based approaches, MoDA adopts a far more compact latent representation—70 channels, compared with Hallo’s $64 \times 64 \times 4$ feature map. As a result, synthesizing 16 frames entails 11,599.69 GFLOPs for Hallo but only 19.53 GFLOPs for MoDA. In addition, we replace the standard DDPM sampler with rectified flow, cutting the number of denoising steps from 50 to 10 while attaining even higher visual quality (see Table 1). These results indicate that MoDA allows us to trim 80 % of the diffusion iterations without compromising visual quality and lip-sync

Component	Latency
Pre-process	0.1303
Denosing Network	0.2012
Rendering	6.3202

Table 2: Single-step inference latency (seconds) for each module.

Method	Lip Sync \uparrow	Motion Diversity \uparrow	ID Similarity \uparrow
EchoMimic	2.6	2.4	3.0
JoyHallo	3.3	2.8	2.9
Hallo	2.8	2.7	3.1
Hallo2	3.2	3.1	3.4
JoyVASA	2.4	1.5	4.4
Ditto	3.1	2.9	4.5
Ours	4.5	4.2	4.1

Table 3: User study with 20 participants (scores range from 1 to 5).

accuracy. Consequently, MoDA delivers state-of-the-art realism at real-time inference speed. A fine-grained runtime profile for generating a 10-second video is reported in Table 2.

Experiments

User Studies

We conducted a user study on the public HDTF dataset, where 20 participants rated the results of six methods on a five-point scale with respect to three key aspects: lip-sync quality, motion diversity, and identity similarity. As reported in Table 3, MoDA achieves the highest scores on all criteria, outperforming every competing approach.

Emotion control

It should be stressed that the perceived emotion in a talking-head sequence is closely coupled with the accompanying audio and can, in most cases, be inferred directly from it. We therefore treat the external emotion cue merely as a "catalyst" that enables MoDA to infer this affect more effectively from the speech signal. The cue is applied only when necessary, to gently amplify or attenuate the latent emotion; it is not designed to perform a full emotion transfer or to synthesize expressions that contradict the input soundtrack. To verify this claim, we reran inference with different emotion codes and measured both lip-sync accuracy and FVD. As shown in Table 4, altering the emotion code does not degrade the model's generation quality.

Additional Ablations and Results

Cross-attention For the cross-attention, we follow Ditto's design (Li et al. 2024): the audio, identity, and emotion embeddings are concatenated along the channel dimension, processed by a four-layer MLP, and then used as the key-value inputs in a cross-attention operation with the noise latent. We also identify an additional weakness of the

Emotion	Sync-C \uparrow	Sync-D \downarrow	FVD \downarrow
Anger	5.885	8.160	207.325
Contempt	5.802	8.210	215.016
Disgust	5.797	8.220	220.178
Fear	5.855	8.131	203.160
Happiness	5.737	8.356	214.524
Neutral	5.804	8.182	206.250
Sadness	5.802	8.227	208.972
Surprise	5.844	8.197	212.952
None (ours)	5.878	8.135	205.442

Table 4: Impact of forcing different emotion labels during inference on lip-sync accuracy and perceptual quality.



Figure 3: Limitations of cross-attention in profile cases

cross-attention baseline in side-view scenarios. As shown in Fig. 3, the head generated with cross-attention rotates toward a frontal pose that does not match the side-view reference. This failure again reflects cross-attention's inability to resolve the inherent one-to-many ambiguity of talking-head generation, causing the model to gravitate toward the frontal orientations prevalent in the training data. By contrast, MoDA exploits the spatial cues in the reference frame to dynamically modulate the audio features, yielding videos whose head poses remain faithful to the intended side view and appear far more natural.

Rectified Flow We conduct ablation studies on the components of the loss function, as summarized in Table 5. Replacement of rectified flow with standard DDPMs leads to performance degradation across all metrics. Using an L1 loss (w/ RF-L1) improves visual quality (lower FVD), but reduces the accuracy of lip synchronization. Given the importance of synchronization in talking head generation, we adopt rectified flow with an L2 loss (Full Model) as the default setting.

Conditional injection method In our model, the conditions of identity and emotion are treated as two additional streams to balance the identity and emotional signals within the audio. To evaluate this design, we retain only the noise and audio streams, and compare two settings: one where emotion and identity conditions are injected via AdaLN ("w/ AdaLN") (Peebles and Xie 2022), and another that omits the emotion and identity conditions, training MMDIT solely on the audio and noisy motion ("w/o E&I").

As shown in Table 5, when the AdaLN method is used, overall performance drops, suggesting that the extraction of identity and emotional cues from the audio is crucial to balance the inter-modal inconsistency.



Figure 4: Examples illustrating the limitations.

Method	FVD ↓	Sync-C ↑	Sync-D ↓
w/ DDPMs	220.291	5.433	8.832
w/ RF-L1	203.371	5.543	8.656
w/o E&I	208.372	5.387	8.619
w/ AdaLN	216.871	5.442	8.441
Full Model	205.442	5.878	8.135

Table 5: Ablation study results on major architectural choices.

Additionally, when identity and emotion conditions are not injected, we observe a significant decline in lip synchronization metrics. This suggests that, without these auxiliary conditions to act as catalysts, it is difficult to rely on the audio alone to generate consistent mouth shapes and natural movements. Interestingly, this variant attains a slightly lower (better) FVD than the AdaLN counterpart, implying that the audio stream already carries rich emotional and identity cues and that the AdaLN strategy cannot fully reconcile the inter-modal inconsistency. Moreover, regardless of whether additional conditions are injected, overall metrics remain superior to those of the cross-attention-based method, further demonstrating the importance of effective information interaction.

Limitations

The main limitation of our framework stems from the first-stage model, where the Liveportrait generator struggles with large pose variations and complex head accessories. As shown in Fig. 4, these issues result in a noticeable decline in output quality, particularly under large pose variations and in the presence of complex head accessories. Significant pose changes often cause unnatural distortions, which are visually disturbing for users. Additionally, complex headwear can be

misinterpreted as part of the background, leading to temporal inconsistencies and blurring between video frames.

References

- Li, T.; Zheng, R.; Yang, M.; Chen, J.; and Yang, M. 2024. Ditto: Motion-Space Diffusion for Controllable Realtime Talking Head Synthesis. *arXiv preprint arXiv:2411.19509*.
- Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers.
- Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-training for Speech Recognition. In *Interspeech 2019*.