TAVID: Text-Driven Audio-Visual Interactive Dialogue Generation

Ji-Hoon Kim[1,⋆], Junseok Ahn[2,⋆], Doyeop Kwak[2], Joon Son Chung[2], and Shinji Watanabe[3]

[1] Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University, Seoul 06974, Republic of Korea
[2] School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea
[3] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. The objective of this paper is to jointly synthesize interactive videos and conversational speech from text and reference images. With the ultimate goal of synthesizing human-like conversations, recent studies have explored talking or listening head generation as well as conversational speech generation. However, these works are typically studied in isolation, overlooking the multimodal nature of human conversation, which involves tightly coupled audio-visual interactions. In this paper, we introduce TAVID, a unified framework that generates both interactive faces and conversational speech in a synchronized manner. TAVID integrates face and speech generation pipelines through two cross-modal mappers (i.e., a motion mapper and a speaker mapper), which enable bidirectional exchange of complementary information between the audio and visual modalities. Extensive experiments demonstrate that TAVID generates realistic and well-synchronized interactive dialogues, achieving strong visual quality, accurate lip synchronization, and natural turn-taking.

1 Introduction

Synthesizing highly realistic, multi-speaker conversational scenes directly from text scripts represents a significant frontier in generative AI. Natural human conversation is an inherently dyadic and multimodal process [9, 22, 29, 35] that requires the holistic synthesis of both the acoustic dialogue and the bidirectional visual interactions. However, prior work has predominantly approached face animation as a single-role task, including talking head [3,7,13,14,24,33] and listening head generation [19, 21, 27, 37], which focus solely on one-sided communication. While recent interactive head generation methods [10,28,30,36,38] model dyadic interactions, they still rely on pre-recorded audio and cannot create new speech content, whereas cascaded text-to-speech (TTS) workarounds suffer from error accumulation [13, 32].

⋆ Equal contribution.

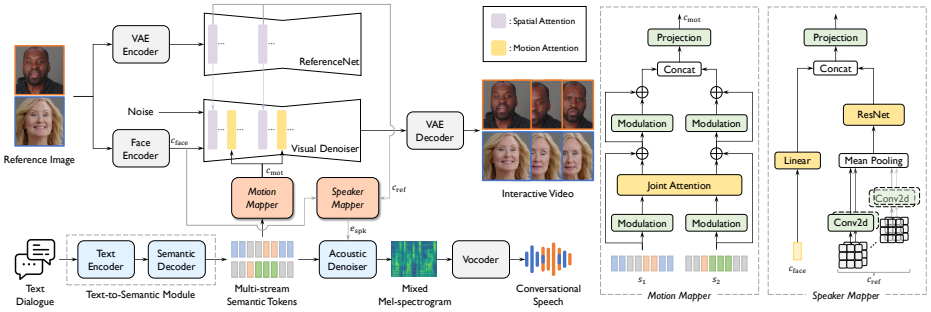

Fig. 1: The overall architecture of TAVID. Given a text dialogue and a reference image, our approach jointly synthesizes interactive video and conversational speech, guided by two cross-modal mappers.

In this paper, we propose TAVID, a unified framework for **T**ext-driven **A**udio-**V**isual **I**nteractive **D**ialogue generation. TAVID jointly generates conversational speech and interactive videos from a text dialogue and reference images through two novel cross-modal mappers, enabling flexible content creation and automatic speaker modeling. The *Motion Mapper* converts the dialogue into dyadic motion features that dynamically alternate between speaking and listening states, adopting a joint self-attention mechanism for bidirectional information exchange between speakers. Meanwhile, the *Speaker Mapper* produces speaker features consistent with the visual identity, ensuring the synthesized voice aligns with the target participant's visual persona. To our knowledge, this is the first attempt to jointly synthesize interactive video and conversational speech from text scripts.

2 TAVID

2.1 Network Architecture

As depicted in Fig. 1, TAVID consists of two key pipelines, interactive video generation and conversational speech synthesis, interconnected through the proposed cross-modal mappers.

Text-to-Semantic Module. Following CoVoMix [34], the text-to-semantic module converts a text dialogue $\mathbf{x}$ into dual-stream semantic tokens $\mathbf{S} = [s_1, s_2]$, one per participant. Distinct from CoVoMix, we adopt the SeamlessM4T [1] semantic tokens, which carry rich paralinguistic information [15]; we hypothesize such cues benefit not only expressive speech [25,26] but also dynamic facial video, as validated in Sec. 3.

Video Generation. Our video pipeline builds on the latent-diffusion talking-head model Hallo2 [5], whose visual denoiser is conditioned on reference spatial features c_{ref} from a parallel ReferenceNet, a face embedding c_{face}, and the interactive motion features c_{mot} produced by the *Motion Mapper* from $\mathbf{S}$—the key

driver of natural dyadic interactions. Both participants are produced by swapping the input streams, using $S = [s_1, s_2]$ for s_1 and the reversed $S' = [s_2, s_1]$ for the interlocutor s_2.

Speech Generation. To generate the conversational speech, a flow-matching [18] acoustic denoiser converts $\mathbf{S}$ into a mel-spectrogram that is rendered by a pre-trained vocoder [16]. Per-stream speaker embeddings condition the voice: audio-driven embeddings [31] are used during training, whereas at inference the *Speaker Mapper* predicts face-driven embeddings from the reference image.

Cross-modal Mappers. Conditioning both pipelines on the shared tokens $\mathbf{S}$ alone is insufficient to capture the rich audio-visual correlations of real conversation. We therefore introduce two cross-modal mappers, described below, that form a bidirectional pathway between the video and speech pipelines to yield tightly coupled audio-visual outputs.

2.2 Motion Mapper

The *Motion Mapper* aims to generate interactive motions from multi-stream semantic tokens $\mathbf{S}$, which requires capturing both the stream-specific information and the interdependencies between the two streams. To this end, we introduce a joint self-attention mechanism, inspired by the MMDiT block [8], which effectively models both modality-specific behaviors and cross-modal correlations between text and image. As shown in Fig. 1, each semantic stream is modulated (LayerNorm and Linear) before a joint self-attention layer captures the mutual dependencies between the two streams; the refined representations are then concatenated and projected by Linear layers into the interactive motion feature c_{mot}, capturing both stream-specific cues and cross-stream correlations.

2.3 Speaker Mapper

To tightly align acoustic and visual identity, the *Speaker Mapper* estimates vocal characteristics from two complementary visual features readily available from the video pipeline: the face embedding c_{face}, which provides the primary identity cue, and the ReferenceNet spatial features c_{ref}, which offer rich discriminative cues [4, 6, 20]. As shown in Fig. 1 (rightmost), c_{ref} and c_{face} are fused through a lightweight network [11] to predict the audio-driven embedding $e_{\text{spk}}^{\text{audio}}$. The *Speaker Mapper* f_θ is trained with an L2 regression loss between its prediction $f_\theta(c_{\text{face}}, c_{\text{ref}})$ and the target audio-driven embedding $e_{\text{spk}}^{\text{audio}}$.

3 Experimental Results

Interactive Head Generation Evaluation. As shown in Fig. 2, TAVID produces accurate lip synchronization, expressive non-verbal motions, and smooth

Table 1: Quality comparison of **interactive head generation** on the Seamless Interaction test set. Subjective evaluation is presented with 95% confidence intervals.

Method	Source	Subjective Metrics			Objective Metrics							
		Visual Quality↑	Lip Sync↑	Turn-taking↑	FID↓	FVD↓	LPIPS↓	LSE-C↑	LSE-D↓	RPCC↓	ΔSID↓	ΔVar↓
DIM [28]	Audio	2.09±0.20	2.00±0.21	2.33±0.23	29.715	279.673	0.240	2.620	11.366	0.086	0.902	0.110
DIM [28]	TTS	2.11±0.23	2.33±0.24	2.38±0.22	30.423	283.056	0.255	2.480	11.620	0.061	1.010	0.229
Ours	Text	**3.75±0.23**	**3.80±0.23**	**3.84±0.24**	**16.625**	**179.305**	**0.056**	**6.457**	**8.403**	**0.031**	**0.489**	**0.011**

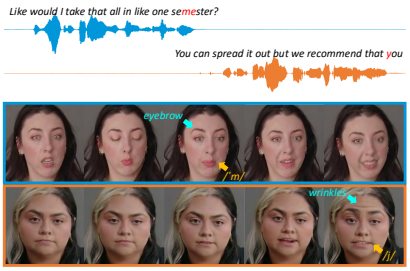

Fig. 2: Qualitative example of TAVID on the Seamless Interaction test set.

Table 2: Ablation on paralinguistic tokens $\mathbf{S}_{\text{para}}$ for talking (THG) and listening (LHG) head generation; bold is better.

Method	THG		LHG			
	FVD↓	LSE-C↑	FD↓		RPCC↓	
			Exp	Pose	Exp	Pose
TAVID	**245.493**	**7.537**	**16.03**	**0.04**	**0.01**	**0.02**
w/o $\mathbf{S}_{\text{para}}$	266.942	6.645	16.39	0.05	**0.01**	0.04

transitions between speaking and listening states, with prosodic emphasis reflected in dynamic facial expressions. Here, emphasized syllables (/'m/, /j/; yellow) drive precise lip motions, while prosodic stress elicits raised eyebrows and forehead wrinkles (cyan). Notably, from text input alone, TAVID generates overlapped video and speech with appropriate, fully synchronized timing.

We compare TAVID against DIM [28], including a TTS-cascaded variant using our generated audio (Tab. 1). TAVID significantly outperforms the baseline across all metrics, and the cascaded DIM degrades further, underscoring the advantage of our end-to-end approach.

Paralinguistics in Video Generation. To identify the effect of paralinguistic information, we compare TAVID (with paralinguistic tokens $\mathbf{S}_{\text{para}}$) against a HuBERT [12]-based variant with limited prosodic cues [2, 17, 23]. As shown in Tab. 2, omitting them degrades both talking and listening head quality—most notably lip-sync and motion coordination (RPCC)—confirming the effectiveness of $\mathbf{S}_{\text{para}}$ for realistic, interactive facial dynamics.

4 Conclusion

In this work, we presented TAVID, the first unified framework to jointly synthesize conversational speech and dyadic visual interactions directly from text scripts through two cross-modal mappers. Extensive experiments confirm its effectiveness in producing realistic interactive dialogues, with paralinguistic tokens further improving expressive facial dynamics. We believe TAVID provides a solid foundation for human-like conversational systems and highlights the value of modeling tightly coupled audio-visual interactions in multimodal synthesis.

References

1. Barrault, L., Chung, Y.A., Meglioli, M.C., Dale, D., Dong, N., Duquenne, P.A., Elsahar, H., Gong, H., Heffernan, K., Hoffman, J., et al.: SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. arXiv:2308.11596 (2023)
2. Chang, H.J., Yang, S.w., Lee, H.y.: DistilHuBERT: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit BERT. In: Proc. ICASSP (2022)
3. Choi, J., Kim, M., Park, S.J., Ro, Y.M.: Text-driven Talking Face Synthesis by Reprogramming Audio-driven Models. In: Proc. ICASSP (2024)
4. Clark, K., Jaini, P.: Text-to-Image Diffusion Models are Zero-Shot Classifiers. In: Proc. NeurIPS (2023)
5. Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., Wang, J.: Hallo2: Long-Duration and High-Resolution Audio-Driven Portrait Image Animation. In: Proc. ICLR (2024)
6. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. In: Proc. NeurIPS (2021)
7. Diao, X., Cheng, M., Barrios, W., Jin, S.: FT2TF: First-Person Statement Text-To-Talking Face Generation. In: Proc. WACV (2025)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: Proc. ICML (2024)
9. Fauviaux, T., Marin, L., Parisi, M., Schmidt, R., Mostafaoui, G.: From unimodal to multimodal dynamics of verbal and nonverbal cues during unstructured conversation. PLOS ONE **19**(9), 1–26 (2024)
10. Guo, Y., Liu, X., Zhen, C., Yan, P., Wei, X.: ARIG: Autoregressive Interactive Head Generation for Real-time Conversations. arXiv:2507.00472 (2025)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. CVPR (2016)
12. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: HuBERT: Self-supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
13. Jang, Y., Kim, J.H., Ahn, J., Kwak, D., Yang, H.S., Ju, Y.C., Kim, I.H., Kim, B.Y., Chung, J.S.: Faces that Speak: Jointly Synthesising Talking Face and Speech from Text. In: Proc. CVPR (2024)
14. Jang, Y., Rho, K., Woo, J., Lee, H., Park, J., Lim, Y., Kim, B.Y., Chung, J.S.: That's What I Said: Fully-Controllable Talking Face Generation. In: Proc. ACM MM (2023)
15. Kim, H., Seo, S., Jeong, K., Kwon, O., Kim, S., Kim, J., Lee, J., Song, E., Oh, M., Ha, J.W., et al.: Paralinguistics-Aware Speech-Empowered Large Language Models for Natural Conversation. In: Proc. NeurIPS (2024)
16. Lee, S.g., Ping, W., Ginsburg, B., Catanzaro, B., Yoon, S.: BigVGAN: A Universal Neural Vocoder with Large-scale Training. In: Proc. ICLR (2023)
17. Lin, G.T., Feng, C.L., Huang, W.P., Tseng, Y., Lin, T.H., Li, C.A., Lee, H.y., Ward, N.G.: On the Utility of Self-supervised Models for Prosody-related Tasks. In: IEEE Spoken Language Technology workshop (2023)
18. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: Proc. ICLR (2023)
19. Liu, X., Guo, Y., Zhen, C., Li, T., Ao, Y., Yan, P.: CustomListener: Text-Guided Responsive Interaction for User-Friendly Listening Head Generation. In: Proc. CVPR (2024)

20. Mukhopadhyay, S., Gwilliam, M., Yamaguchi, Y., Agarwal, V., Padmanabhan, N., Swaminathan, A., Zhou, T., Ohya, J., Shrivastava, A.: Do Text-free Diffusion Models Learn Discriminative Visual Representations? In: Proc. ECCV (2024)
21. Ng, E., Joo, H., Hu, L., Li, H., Darrell, T., Kanazawa, A., Ginosar, S.: Learning to listen: Modeling non-deterministic dyadic facial motion. In: Proc. CVPR (2022)
22. Nota, N., Trujillo, J.P., Holler, J.: Facial signals and social actions in multimodal face-to-face interaction. Brain Sciences **11**(8), 1017 (2021)
23. Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.N., Mohamed, A., Dupoux, E.: Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In: Proc. Interspeech (2021)
24. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In: Proc. ACM MM (2020)
25. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y.: FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In: Proc. ICLR (2021)
26. Ren, Y., Lei, M., Huang, Z., Zhang, S., Chen, Q., Yan, Z., Zhao, Z.: ProsoSpeech: Enhancing Prosody With Quantized Vector Pre-training in Text-to-Speech. In: Proc. ICASSP (2022)
27. Song, L., Yin, G., Jin, Z., Dong, X., Xu, C.: Emotional Listener Portrait: Realistic Listener Motion Simulation in Conversation. In: Proc. ICCV (2023)
28. Tran, M., Chang, D., Siniukov, M., Soleymani, M.: DIM: Dyadic Interaction Modeling for Social Behavior Generation. In: Proc. ECCV (2024)
29. Truong, K.P., Poppe, R., Kok, I.d., Heylen, D.: A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In: Proc. Interspeech (2011)
30. Wang, D., Dai, B., Deng, Y., Wang, B.: Disentangling Planning, Driving and Rendering for Photorealistic Avatar Agents. In: Proc. ECCV (2024)
31. Wang, H., Zheng, S., Chen, Y., Cheng, L., Chen, Q.: CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking. In: Proc. Interspeech (2023)
32. Wang, Z., Zhang, P., Qi, J., Xu, G.W.S., Zhang, B., Bo, L.: OmniTalker: Real-Time Text-Driven Talking Head Generation with In-Context Audio-Visual Style Replication. arXiv:2504.02433 (2025)
33. Xu, S., Chen, G., Guo, Y.X., Yang, J., Li, C., Zang, Z., Zhang, Y., Tong, X., Guo, B.: VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time. In: Proc. NeurIPS (2024)
34. Zhang, L., Qian, Y., Zhou, L., Liu, S., Wang, D., Wang, X., Yousefi, M., Qian, Y., Li, J., He, L., et al.: CoVoMix: Advancing Zero-Shot Speech Generation for Human-like Multi-talker Conversations. In: Proc. NeurIPS (2024)
35. Zhang, Y., Frassinelli, D., Tuomainen, J., Skipper, J.I., Vigliocco, G.: More than words: word predictability, prosody, gesture and mouth movements in natural language comprehension. Proceedings of the Royal Society B: Biological Sciences **288**(1955), 20210500 (2021)
36. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T.: Interactive Conversational Head Generation. IEEE TPAMI (2025)
37. Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., Mei, T.: Responsive Listening Head Generation: A Benchmark Dataset and Baseline. In: Proc. ECCV (2022)
38. Zhu, Y., Zhang, L., Rong, Z., Hu, T., Liang, S., Ge, Z.: INFP: Audio-Driven Interactive Head Generation in Dyadic Conversations. In: Proc. CVPR (2025)