UniLS: End-to-End Audio-Driven Avatars for Unified Listening and Speaking

Supplementary Material

7. Evaluation of audio-free generator

We evaluate the motion produced by the audio-free generator trained in stage 1. During evaluation, the stage-1 model is tested in a free autoregressive manner, where only the initial motion chunk is provided and all subsequent motions are generated without any audio guidance. As shown in Tab. 5, we compare its outputs with the full two-stage UniLS model across both speaking and listening metrics. Although the audio-free generator receives no audio input, it still produces motions with noticeable diversity and natural temporal variation, particularly in metrics related to dynamic deviation (FDD, PDD, JDD). This shows that stage 1 effectively captures human's internal motion priors, which prevents the model from collapsing into trivial or frozen outputs and instead encourages natural variability.

8. User study details

Fig. 6 shows the interface of our user study. We collect responses from 25 participants, each completing 32 pairwise comparison trials (128 questions in total). All comparisons are performed on videos from the Seamless Interaction dataset. To avoid positional bias, the assignment of our method and the baseline to "A" or "B" is randomized for every trial. Each comparison includes four single-choice questions. Following prior works [11, 40], the first two assess lip synchronization and expression naturalness. We further introduce two questions targeting conversational behavior: reaction naturalness and head pose naturalness. For each question, participants select which video (A or B) appears more natural or better aligned with the audio.

Table 5. Quantitative evaluation of the audio-free generator trained in stage 1.

Method	Speak					Listen				
	LVE↓	MHD↓	FDD↓	PDD↓	JDD↓	FDD↓	PDD↓	JDD↓	F-FID↓	P-FID↓
Audio-free generator (stage 1 only)	-	-	23.97	5.61	1.34	21.83	5.17	1.59	6.45	0.053
UniLS (stage 1 + stage 2)	5.83	1.89	18.41	4.67	0.71	17.12	4.75	0.98	4.30	0.038

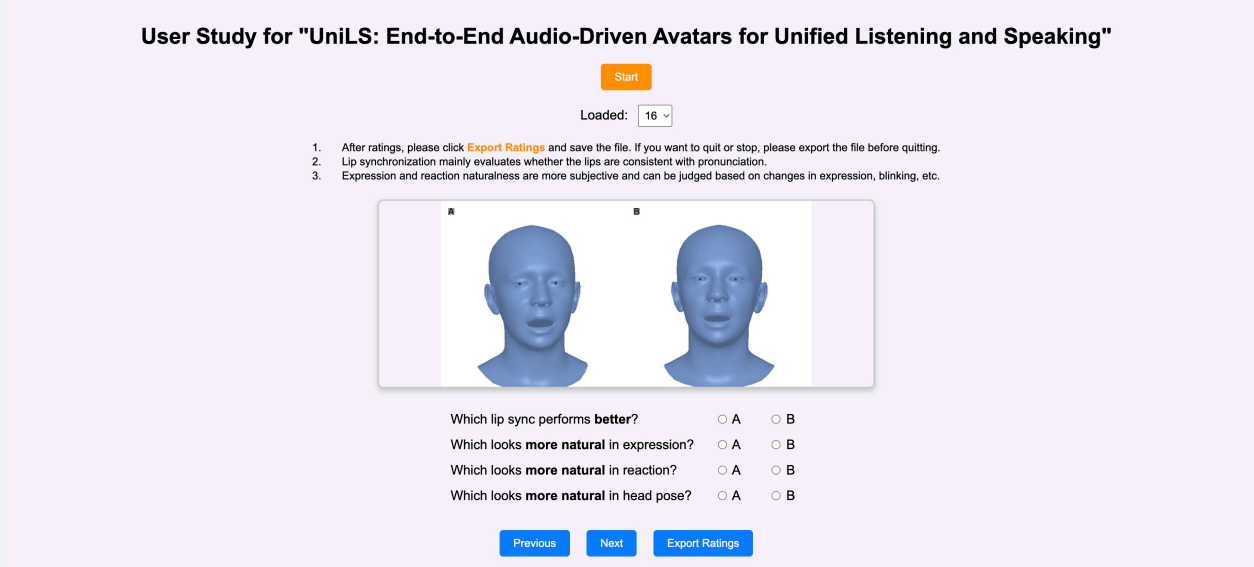

Figure 6. The interface of our user study. Users evaluate each video based on four perspectives: lip synchronization, expression naturalness, reaction naturalness, and head pose naturalness. All these four perspectives are judged by comparing methods A and B. One of the videos (A or B) is generated by our method, and the other by a baseline method, with their order randomized.

9. Ethics statement

UniLS is designed to generate audio-driven 3D conversational avatars. While such technology has clear benefits in domains like education and human–computer interaction, it also presents potential ethical risks if misused, such as impersonation, generating misleading content, or producing avatars of real individuals without consent. We are aware of these concerns and strongly discourage any form of misuse. To mitigate these risks, we adopt the following safeguards:

- **Visible watermarking.** We add a clear watermark to all synthesized videos so that viewers can immediately identify them as generated content, reducing the risk of deception.
- **Restricted identity usage.** We limit the target identities to virtual characters (*e.g.*, virtual idols) and prohibit synthesizing real individuals without formal consent. Any generated content may only be used for educational or other legitimate purposes, and misuse is subject to accountability as described below.
- **Invisible watermarking and provenance.** We inject invisible watermarks into synthesized videos to record the source (*e.g.*, the producer's IP), ensuring traceability and encouraging creators to carefully consider ethical implications before generating content.

In summary, we provide strict licensing terms and technical measures to minimize the risk of abuse of UniLS. Broader efforts from researchers, practitioners, policymakers, and users are needed to ensure responsible development and use of avatar-generation technologies. With appropriate safeguards and ethical use, UniLS offers meaningful benefits for a wide range of real-world applications.