UniAVGen: Unified Audio and Video Generation with Asymmetric Cross-Modal Interactions

Supplementary Material

1. Additional implementation details

1.1. Context window size for A2V aligner

In the Asymmetric Cross-Modal Interaction mechanism (Sec. 3.2 of the main paper), the audio context window size w for A2V aligner is set to $\frac{1}{2}$. Specifically, for the i-th video latent frame, the audio context window C_i^a concatenates audio tokens from $i-\frac{1}{2}$ to $i+\frac{1}{2}$ (i.e., 2 audio segments in total), ensuring sufficient contextual phoneme information for precise lip synchronization. Boundary frames (when $i-\frac{1}{2}<0$ or $i+\frac{1}{2}\geq T$) are padded by replicating the first or last audio frame's features to avoid information loss.

1.2. Temporal alignment in interaction

Due to the design of existing video VAEs, each video latent, except the first one which corresponds to a single frame, is associated with four consecutive frames. To ensure precise temporal alignment during cross-modal interaction, we explicitly account for this characteristic of video latents. Specifically, for A2V alignment, we first compute the audio window size per frame by dividing the number of audio tokens by the actual number of video frames. We then determine the corresponding audio window for each video latent, meaning the effective window for the first latent is a quarter the size of those for subsequent latents. For V2A alignment, we first upsample the video latents to match audio's fine-grained temporal resolution: each latent (except the first) is replicated four times. With this temporal alignment, we then compute the video context.

1.3. Inference details

We employ the Euler ODE solver with 50 sampling steps and leverage the Vocos vocoder [10] to convert generated log mel spectrograms into audio signals. For MA-CFG, we empirically set $s_v = 3$ and $s_a = 2$. To stabilize audio-visual quality while using CFG, we further adopt CFG interval [7], which restricts classifier-free guidance exclusively to the high-frequency generation phase with the interval set to $[0.5, 1]$. Additionally, for efficiency, we set the text condition to empty during unimodal sampling in MA-CFG, which further reinforces text control.

2. System prompt for evaluation

We use the following system prompts to evaluate Timbre Consistency (TC) and Emotion Consistency (EC) via Gemini-2.5-Pro. The prompt is designed to ensure objective, reproducible scoring (0-1 scale, 2 decimal places):

```
You are an expert in audio and video understanding.
Now you will receive an audio and video clip. Plea-
se judge the consistency between the timbre and em-
otion of the audio and video, and give a score bet-
ween 0 and 1.

For timbre evaluation (score a), it is divided into
5 grades based on gender and age matching:
1. 0 points: Completely inconsistent (e.g., video
shows a woman but audio is a man's voice; age diff-
erence is extremely obvious)
2. 0.25 points: Severely inconsistent (one of gende-
r or age is seriously mismatched, the other has sli-
ght inconsistency)
3. 0.5 points: Partially inconsistent (one of gender
or age is mismatched, the other is consistent)
4. 0.75 points: Basically consistent (gender and age
are roughly matched, with minor details inconsistent)
5. 1 point: Perfectly consistent (gender and age are
completely matched without any differences)

For emotion evaluation (score b), it is divided into
5 grades based on frame-level emotion matching and
body language correspondence:
1. 0 points: No correspondence at all (no frame mat-
ches, body language has nothing to do with audio)
2. 0.25 points: Rarely corresponding (very few frames
match, body language basically does not correspond)
3. 0.5 points: Partially corresponding (about half of
the frames match, body language partially corresponds)
4. 0.75 points: Basically corresponding (most frames
match, body language roughly corresponds)
5. 1 point: Perfectly corresponding (every frame mat-
ches, body language fits audio perfectly)

You should return the following JSON format:

{"score":[a, b],"reason":"xxx"}

Where a is the timbre score, b is the emotion score,
and reason is the specific reason for the score, wh-
ich should not exceed 100 words.
```

Each sample is evaluated 3 times independently, and the average score is reported.

Table 1. **User study statistics.**

Methods	AQ(↑)	VQ(↑)	AVC(↑)
Universe-1 [11]	2.35%	0.00%	0.00%
Ovi [8]	28.75%	37.40%	25.70%
UniAVGen (ours)	**68.90%**	**62.60%**	**74.30%**

3. User study

A comprehensive user study was also performed to further underscore the advantages of our method. Participants evaluated and selected the top-generated videos by assessing

Table 2. **Results on GRID [4] under Dubbing Setting 3.0.**

Methods	LSE-C(↑)	LSE-D(↓)	WER(↓)
StyleDubber [2]	5.94	9.75	15.40
EmoDubber [3]	7.25	6.83	14.72
UniAVGen (ours)	**7.59**	**6.11**	**10.64**

Table 3. **Results on EMTD [9].**

Methods	LSE-C(↑)	LSE-D(↓)	FID(↓)	FVD(↓)
OmniAvatar [5]	7.19	6.90	45.02	459.44
Wan-S2V [6]	**7.24**	6.92	44.02	**451.44**
UniAVGen (ours)	7.05	**6.85**	**43.97**	469.85

audio quality (AQ), video quality (VQ), and overall audio-visual coherence (AVC). Results from 34 participants, presented in Tab. 1, reveal that our approach achieves superior overall audio-visual quality and enhanced consistency between audio and video compared to recent methods.

4. Evaluation on conditional tasks

While UniAVGen is primarily designed for high-quality joint audio-visual generation, we further evaluate its performance on public benchmarks of other conditional generation tasks after multi-task joint training to ensure the completeness of this work. For video-to-audio dubbing, we test on the widely used GRID benchmark [4] with three metrics: LSE-C [1], LSE-D [1] and WER. As shown in Tab. 2, we compare performance under Dubbing Setting 3.0, which adopts unseen speakers as reference audio. Without complex or task-specific designs, our model achieves superior consistency and lower WER. For audio-to-video synthesis, we utilize the half-body animation benchmark EMTD [9] and compare against state-of-the-art audio-driven models [5, 6]. As presented in Tab. 3, our model attains near-SOTA performance with only simple multi-task fine-tuning. These results further validate the practicality and generalization capability of UniAVGen.

5. Extended ablation studies

We supplement additional ablation experiments to validate the robustness of our core designs:

5.1. Exploration of interaction insertion positions

Rationally integrating the interaction module is another critical consideration, which we address from two perspectives. First, at the layer-level (see Tab. 4), we explore four schemes: inserting into all layers, the first half of layers, the last half of layers, and interleaved insertion. Interleaved insertion yields the best results, indicating that appropriate yet not excessive cross-modal interaction better enhances the stability of multi-modal learning. Second, at the operation-level: built on the DiT architecture of Wan2.2-5b, each DiT block comprises self-attention, text cross-attention, and FFN. We ablate the module insertion at three distinct positions: 1) before self-attention, 2) before cross-attention, and 3) before FFN. As shown in Tab. 5, position 1) achieves the optimal performance, suggesting that fully preserving the operational flow of each block facilitates better inheritance of pretrained capabilities.

Table 4. **Ablation studies on the layer-level insertion.**

Settings	LS(↑)	TC(↑)	EC(↑)
(a) all layers	4.01	0.713	0.497
(b) first half of layers	4.02	0.719	0.500
(c) last half of layers	3.79	0.710	0.493
(d) interleaved layers	**4.09**	**0.725**	**0.504**

Table 5. **Ablation studies on the operation-level insertion.**

Settings	LS(↑)	TC(↑)	EC(↑)
1) before FFN	3.85	0.715	0.490
2) before cross-attention	3.98	0.721	0.499
3) before self-attention	**4.09**	**0.725**	**0.504**

Table 6. **Ablation studies on the MA-CFG.**

Settings	LS(↑)	TC(↑)	EC(↑)	IQ(↑)
(a) no CFG	5.75	0.821	0.553	0.760
(b) vanilla CFG	5.81	0.824	0.562	0.778
(c) MA-CFG	**6.29**	**0.841**	**0.580**	0.752
(d) MA-CFG under $t \in [0.5, 1]$	5.95	0.832	0.573	**0.779**

5.2. Validation of MA-CFG's effectiveness

As shown in Tab. 6, we compare the performance of four testing strategies: no CFG, vanilla CFG, MA-CFG, and MA-CFG with the interval $[0.5, 1]$. While vanilla CFG improves image quality, its enhancement on modal consistency is negligible. In contrast, MA-CFG significantly boosts audio-visual alignment metrics but slightly degrades image quality. By incorporating the constrained CFG interval, MA-CFG achieves simultaneous improvements in both image quality and modal alignment.

6. Limitations

Currently, while UniAVGen performs well in speech-video generation, it lacks video-aligned ambient sound generation. Additionally, its ability to generate audio for multi-person scenarios remains constrained by the inflexible text encoder. For future efforts, we will first collecting more general high-quality audio-video data. Meanwhile, we plan to enhance the text encoder of the audio branch, specifically by adopting multi-modal large language models like Qwen-Omni3 [12], to enable multi-person scenarios generation.

References

[1] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2

[2] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. Styledubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6767–6779, 2024. 2

[3] Gaoxiang Cong, Jiadong Pan, Liang Li, Yuankai Qi, Yuxin Peng, Anton van den Hengel, Jian Yang, and Qingming Huang. Emodubber: Towards high quality and emotion controllable movie dubbing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15863–15873, 2025. 2

[4] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. 2

[5] Qijun Gan, Ruizi Yang, Jianke Zhu, Shaofei Xue, and Steven Hoi. Omniavatar: Efficient audio-driven avatar video generation with adaptive body animation. *arXiv preprint arXiv:2506.18866*, 2025. 2

[6] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025. 2

[7] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024. 1

[8] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. 1

[9] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5489–5498, 2025. 2

[10] Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *arXiv preprint arXiv:2306.00814*, 2023. 1

[11] Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025. 1

[12] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2