Harmony: Harmonizing Audio and Video Generation through Cross-Task Synergy (Supplementary Material)

Teng Hu[1*] Zhentao Yu[2*] Guozhen Zhang[2] Zihan Su[1] Zhengguang Zhou[2]
Youliang Zhang[2] Yuan Zhou[2] Qinglin Lu[2] Ran Yi[1†]
[1]Shanghai Jiao Tong University [2]Tencent Hunyuan

Project page: https://sjtuplayer.github.io/projects/Harmony

1. Overview

In this supplementary material, we provide more implementation details, experiment results, including:

- Implementation details (Sec. 2);
- Benchmark settings (Sec. 3);
- More quantitative comparisons (Sec. 4);
- More qualitative comparisons (Sec. 5);
- Visualization of Cross-modal Attention (Sec. 6)
- More ablation study (Sec. 7);
- Analysis of off-screen audio and SyncCFG behavior (Sec. 8);
- Model agnosticism evaluation (Sec. 9);
- Details about voice clone (Sec. 10);
- Audio-driven performance; (Sec. 11);
- More qualitative results (Sec. 12).

2. Implementation Details

Datasets. Our training corpus is curated from a diverse range of public and newly collected sources to cover both human speech and environmental sounds.

1) Human Speech Data: We aggregate vocal data from multiple open-source datasets, including the TTS-specific Emilia dataset [7], as well as audio-visual corpora such as OpenHumanVid [12] and SpeakerVid [24]. To ensure high-quality alignment, we employed an audio-visual consistency scoring model to filter this collection, resulting in a high-quality subset of 2 million video clips, each 3-10 seconds in duration. We then utilized the Gemini [6] for automated annotation, generating ASR transcripts, descriptive video captions, and captions for any background sounds present in the clips.

2) Environmental Sound Data: For environmental sounds, we leverage several established public datasets, including AudioCaps [11] (∼128 hours, manually captioned), Clotho [4] (∼31 hours, manually captioned), and WavCaps [15] (∼7,600 hours, automatically captioned). Recognizing the often-suboptimal visual quality of the VGGSound dataset [1], we supplemented our data by collecting an additional 2 million audio-visual clips rich in environmental sounds. These new clips were subsequently annotated using Gemini [6] to generate corresponding audio and video captions.

Training Strategy. Our training protocol is structured in three distinct stages to ensure stable convergence and high-fidelity generation. For the video branch, we initialize our model with the pre-trained weights of Wan2.2-5B [22]. The audio model undergoes a dedicated two-stage pre-training process before the final joint training.

Stage 1: Foundational Audio Pre-training. The audio model is first pre-trained on a balanced 1:1 mixture of our human speech and environmental sound datasets. We train for 100,000 iterations with a global batch size of 1536, using clips with a maximum duration of 10 seconds. During this stage, the reference audio is a randomly selected 1-3 second segment from the ground-truth clip. This phase enables the model to learn to replicate both the timbre and content from the provided reference audio.

Stage 2: Timbre Disentanglement Finetuning. To enable the model to disentangle general acoustic characteristics from specific content, we finetune it using mismatched reference and target content. For human speech, we use cross-utterance data from the same speaker. For environmental sounds, we sample a non-overlapping reference clip from the same long recording as the ground-truth target. This setup compels the model to extract the invariant acoustic signature—be it a speaker's voice or an environmental ambience—from the reference and apply it to the new content dictated by the prompt or transcript. We finetune for an additional 20,000 iterations in this configuration.

Stage 3: Cross-Task Audio-Visual Training. Finally, we proceed to the Cross-Task joint training stage. The full audio-visual model is trained for 10,000 iterations with a batch size of 128, again using a 1:1 mixture of human speech and environmental sound data. Across all training

*Equal Contribution
†Corresponding author.

stages, we employ a constant learning rate of 1e-5 for all model parameters.

Hyperparameters. During the final cross-task training stage, the balancing weights for our synergistic loss (Eq. 3) are set to $\lambda_v = 0.1$ and $\lambda_a = 0.3$. The model is trained using a Flow Matching objective with shift of 5. For inference, we use 40 integration steps with classifier-free guidance (CFG) scales of $s_v = 3$ for video and $s_a = 2$ for audio. The sampler's shift parameter is also maintained at 5.

3. Benchmark Settings

3.1. The Harmony-Bench Dataset

Existing benchmarks for audio-visual generation are inadequate for comprehensive evaluation. JavisBench [13] lacks evaluation for human speech, while Verse-Bench [23] is hampered by low-quality labels and a limited focus on audio-visual synchronization. To enable a more rigorous and holistic assessment, we construct and introduce **Harmony-Bench**. This new benchmark features 150 meticulously designed test cases, organized into three progressively challenging subsets (50 items each). It is specifically crafted to disentangle and systematically evaluate a model's semantic consistency and temporal synchronization across diverse and complex acoustic scenarios.

- **Ambient Sound-Video Generation.** This subset is designed to assess the model's ability to generate non-speech acoustic events that are precisely synchronized with corresponding visual dynamics. The 50 test cases feature synthetically constructed scenarios, enabling the creation of complex audio-visual interactions that are difficult to capture or isolate in real-world recordings. The model is conditioned on a detailed `audio_caption` and a separate `video_caption`. Evaluation centers on audio fidelity, temporal synchrony, and the semantic consistency between the generated audio and visual events.
- **Speech-Video Generation.** This 50-item subset assesses the fidelity of speech synthesis and lip synchronization. To test for robustness and multilingual generalization, it includes a balanced mix of 25 real-world and 25 AI-synthesized samples, driven by transcripts in both English (`spoken_word_en`) and Chinese (`spoken_word_zh`). The `video_caption` is deliberately kept minimal (e.g., "a man is speaking"), compelling the model to derive lip movements and facial expressions directly from the transcript's content. Key evaluation criteria are speech intelligibility, naturalness, and the precision of lip-audio synchronization.
- **Complex Scene: Ambient + Speech.** Representing the most challenging scenario, this subset evaluates the model's capacity to simultaneously generate and synchronize both speech and ambient sounds within a unified, complex scene. Each of the 50 test cases is constructed to feature co-occurring audio-visual events, requiring the model to process a combination of inputs: a transcript (`spoken_word_en`), an ambient sound description (`audio_caption`), and a visual scene description (`video_caption`). The evaluation critically examines the model's ability for sound source separation and mixing (e.g., maintaining speech clarity over a background door-closing sound). Furthermore, it assesses multi-modal temporal alignment: speech must synchronize with lip movements, while ambient sounds must align with their corresponding visual actions.

To provide a comprehensive evaluation on this benchmark, we adopt a suite of automated metrics designed to assess three key aspects: 1) Visual Quality and Coherence, 2) Audio Fidelity, and 3) Audio-Visual Synchronization and Consistency.

3.2. Evaluation Metrics

To comprehensively assess model performance on Harmony-Bench, we employ a suite of automated metrics targeting three core aspects of audio-visual quality.

Visual Quality and Coherence. We evaluate the visual quality and temporal consistency of the generated videos using the following metrics:

- **Aesthetic and Imaging Quality.** We assess **aesthetic quality (AQ)** and **imaging quality (IQ)** using the pre-trained aesthetic-predictor-v2-5[21] and MUSIQ[10] models, respectively.
- **Motion Dynamics.** Temporal coherence is evaluated through **Dynamic Degree (DD)** and **Motion Smoothness (MS)**[8]. We employ RAFT[19] to quantify the magnitude of motion and a pre-trained video frame interpolation model to evaluate motion smoothness.
- **Identity Consistency (ID).** For subject-specific generation, we measure ID by computing the mean DINOv3[18] feature similarity between a reference image and all generated frames.
- **CLIP Score.** This metric measures the semantic alignment between the generated video content and the input text prompt.

Audio Fidelity and Quality. The quality of the generated audio is measured by:

- **AudioBox-Aesthetics.**[20] We employ this model to evaluate perceptual quality across four dimensions: **Production Quality (PQ)**, **Production Complexity (PC)**, **Content Enjoyment (CE)**, and **Content Usefulness (CU)**.
- **Word Error Rate (WER).** For speech synthesis, accuracy is measured by WER. We transcribe the generated audio using Whisper-large-v3[16] and compare it against the ground-truth transcript.
- **IB-A Score.** Semantic alignment between the generated audio and the text prompt is quantified using the IB-A

Table 1. **Environment set comparison** with state-of-the-art joint audio-visual generation models. We evaluate performance across three categories: video quality, audio fidelity, and audio-visual synchronization. Best results are in **bold**, second-best are underlined.

Method	Video Quality & Coherence						Audio Fidelity & Quality						Audio-Visual Synchronization			
	AQ ↑	IQ ↑	DD ↑	MS ↑	ID ↑	CLIP ↑	PQ ↑	PC ↓	CE ↑	CU ↑	WER ↓	IB-A ↑	Sync-C ↑	Sync-D ↓	DeSync ↓	IB ↑
MM-Diffusion [17]	0.32	0.43	0.13	0.99	-	-	5.37	4.07	**4.27**	5.89	-	-	-	-	-	0.12
JavisDiT [13]	0.37	0.55	0.33	0.99	0.45	**0.66**	5.64	**2.29**	3.06	5.14	-	0.18	-	-	0.94	0.16
UniVerse-1 [23]	0.57	**0.68**	0.16	**1.00**	0.92	**0.66**	6.14	2.30	3.20	5.46	-	0.04	-	-	1.10	0.07
Ovi [14]	0.62	0.66	0.44	0.99	**0.93**	**0.66**	6.45	2.46	3.78	5.98	-	**0.20**	-	-	1.06	0.20
Harmony (Ours)	**0.64**	0.65	**0.56**	0.98	0.90	**0.66**	**6.53**	2.68	4.12	**6.22**	-	0.14	-	-	**0.70**	**0.21**

Table 2. **Complex set comparison** with state-of-the-art joint audio-visual generation models. We evaluate performance across three categories: video quality, audio fidelity, and audio-visual synchronization. Best results are in **bold**, second-best are underlined.

Method	Video Quality & Coherence						Audio Fidelity & Quality						Audio-Visual Synchronization			
	AQ ↑	IQ ↑	DD ↑	MS ↑	ID ↑	CLIP ↑	PQ ↑	PC ↓	CE ↑	CU ↑	WER ↓	IB-A ↑	Sync-C ↑	Sync-D ↓	DeSync ↓	IB ↑
MM-Diffusion [17]	0.32	0.43	0.13	0.99	-	-	5.37	4.07	4.27	**5.89**	-	-	-	-	-	0.12
JavisDit [13]	0.34	0.50	**0.54**	0.98	0.33	**0.65**	5.40	2.26	2.91	4.56	1.00	**0.09**	0.58	10.50	1.32	0.17
UniVerse-1 [23]	0.52	0.65	0.42	0.99	0.85	**0.65**	6.14	2.23	3.85	5.15	0.25	0.00	0.72	8.32	**1.09**	0.14
Ovi [14]	0.60	0.63	0.41	0.99	0.88	**0.65**	5.94	2.33	4.14	5.33	0.79	0.06	2.94	8.86	1.21	**0.18**
Harmony (Ours)	**0.64**	**0.66**	0.32	**1.00**	**0.91**	**0.65**	**6.43**	**1.90**	**4.76**	4.86	**0.15**	0.06	**4.70**	**6.43**	1.13	**0.18**

Table 3. **Chinese speech comparison** with state-of-the-art models, focusing on audio fidelity (WER) and audio-visual synchronization. Best results are in **bold**, second-best are underlined.

Method	WER ↓	Sync-C ↑	Sync-D ↓	IB ↑
JavisDiT [13]	4.84	1.27	12.63	0.20
UniVerse-1 [23]	2.32	0.91	11.02	**0.22**
Ovi [14]	9.10	4.45	10.79	0.20
Harmony (Ours)	**0.92**	**5.05**	**9.38**	**0.22**

Score[5].

Audio-Visual Synchronization. The critical capability of joint generation is assessed through synchronization metrics:

- **Sync-C & Sync-D.** Lip-sync accuracy is explicitly measured using these two established metrics[3].
- **DeSync Score.** Predicted by Synchformer[9], this score quantifies the temporal misalignment (in seconds) between the audio and video streams.
- **ImageBind (IB) Score.** Following [5], we use the IB score to assess overall audio-visual consistency by computing the cosine similarity between their respective feature embeddings.

4. More Quantitative Comparisons

In this section, we present detailed quantitative comparisons against state-of-the-art methods for joint audio-video generation, including Ovi [14], UniVerse-1 [23], JavisDiT [13], and MM-Diffusion [17]. Our evaluation spans multiple challenging test sets, with results for environmental sounds and complex audio scenes presented in Tables 1–2. Across these diverse datasets, our model consistently demonstrates superior performance. A key observation is our model's superior video dynamism compared to competitors. For instance, while UniVerse-1 and Ovi sometimes achieves a favorable Identity Distance (ID) score, this is often a consequence of generating static or nearly static videos, where frame-to-frame identity is trivially high but fails to capture the scene's intended motion. Crucially, our method consistently achieves the lowest Word Error Rate (WER) and the best scores on audio-visual synchronization metrics. This combination of high fidelity, strong dynamism, and precise alignment underscores our model's robustness in generating coherent and realistic content for complex scenes.

Furthermore, we specifically assess the cross-lingual capabilities of the models on a dedicated Chinese speech test set, with key results summarized in Table 3. The results highlight a significant performance gap. Our model achieves a substantially lower WER and markedly better synchronization scores. It is worth noting that the standard WER metric is not perfectly optimized for the tokenization of the Chinese language; therefore, the relative performance between models serves as the most meaningful indicator. The pronounced improvement in both WER and synchronization metrics strongly validates the effectiveness and superiority of our approach for cross-lingual audio-visual speech generation.

5. More Qualitative Comparisons

In this section, we present further qualitative comparisons of our method against state-of-the-art approaches: Ovi [14], UniVerse-1 [23], and JavisDiT [13]. We focus on two challenging scenarios: synchronized human speech

Figure 1. More comparison on human-speech video generation.

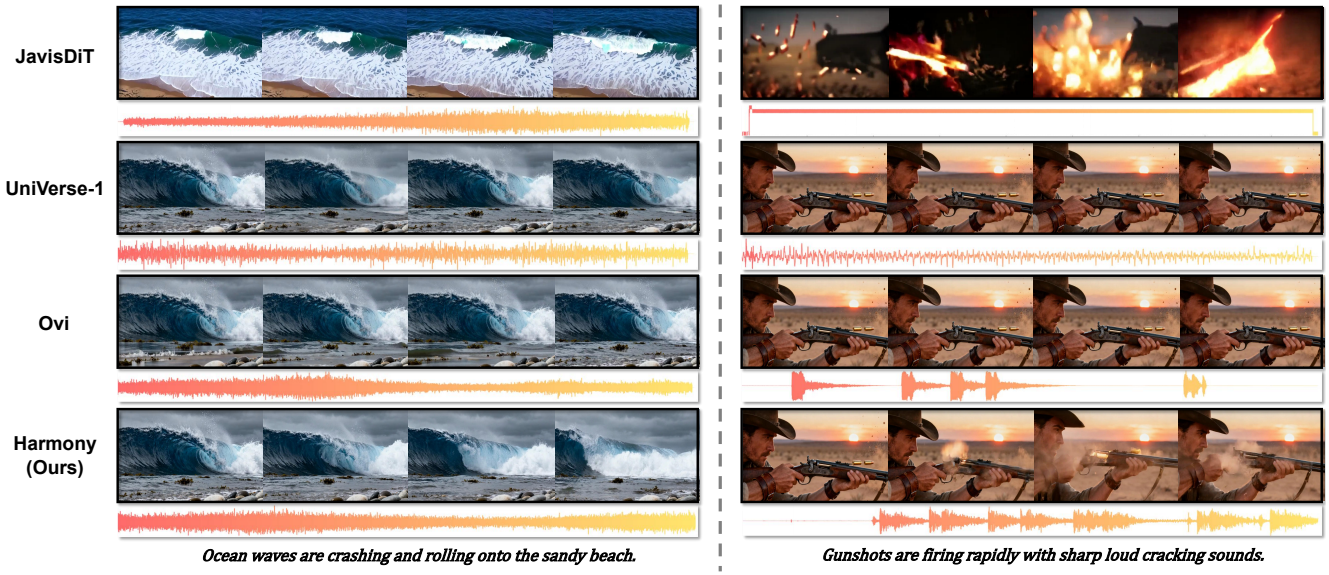

Figure 2. More comparison on environment-sound video generation.

and dynamic environmental sounds. We exclude MM-Diffusion [17] from this analysis as it is designed for unconditional generation and is therefore not directly comparable.

Comparisons on Human Speech. As illustrated in Figure 1, our model demonstrates superior performance in audio-visual speech generation. Competing methods like Ovi and UniVerse-1 tend to produce static or minimally dynamic video frames, resulting in a "talking head" effect with little natural movement. In contrast, our model generates high-fidelity video with fluid, naturalistic motion. The accompanying audio is clear and, most importantly, precisely synchronized with the lip movements, resulting in a significantly more coherent and believable output.

Comparisons on Environmental Sounds. We further evaluate performance on generating dynamic environmental sounds in Figure 2, where the shortcomings of other methods are even more pronounced. JavisDiT struggles in this domain, producing low-quality video and unstable audio; for instance, in the "gunfire" example, its generated audio waveform is highly irregular and fails to represent the acoustic event convincingly. UniVerse-1 and Ovi frequently generate static or partially static scenes. A clear example is the "ocean waves" case, where the main waves remain frozen while only the water surface shows minimal

Figure 3. Visualization of the audio-to-video frame-wise cross-attention map, where the audio can accurately capture the sound source from the videos.

movement. This lack of dynamism is compounded by poor audio-visual synchronization, where the sound of crashing waves does not align with the visual content. In stark contrast, our method excels in all aspects: it generates high-quality, dynamic videos with realistic motion, and the synthesized audio is both high-fidelity and precisely synchronized with the visual events, delivering a cohesive and immersive audio-visual experience.

6. Visualization of Cross-modal Attention

To validate the effectiveness of our frame-wise cross-attention mechanism, we visualize the attention maps from the audio-to-video module. As illustrated in Fig. 3, when synthesizing human speech, the model precisely localizes its attention on the speaker's oral region. Notably, in scenarios with multiple individuals, our model can distinguish between them, focusing exclusively on the active speaker. This capability extends to natural sounds, where the model accurately identifies the primary sound source (e.g., an animal) while also attending to ambient environmental sounds, such as the rain in the cat example and the birdsong in the crocodile case. Collectively, these visualizations underscore our model's superior ability to achieve fine-grained and contextually aware audio-visual alignment.

7. More Ablation Study

In the main paper , we provided an additive ablation study demonstrating the progressive performance improvements brought by each component of the Harmony framework. To further validate the individual necessity of these modules, we present a complementary leave-one-out (subtractive) ablation analysis in this section. We evaluate the models on the human-speech dataset using the Sync-C metric to measure audio-visual synchronization. The quantitative results are summarized in Table 4.

The Dependency of SyncCFG on CTS. It is crucial to note that the proposed Synchronization-Enhanced CFG (SyncCFG) is fundamentally dependent on the Cross-Task Synergy (CTS) training paradigm. SyncCFG operates by explicitly isolating and amplifying the alignment signal using specific "driven" predictions (i.e., mute audio and static video anchors). The capability to generate these meaningful negative anchors is learned exclusively through the auxiliary tasks introduced during CTS training.

As shown in Table 4, if we attempt to apply SyncCFG on a model trained without CTS (*w/o CTS (w/ SyncCFG)*), the performance collapses to a Sync-C score of 1.13. This severe degradation occurs because the model never learned the negative "driven" branch, resulting in significant artifacts during inference. By contrast, a standard baseline without both CTS and SyncCFG achieves a Sync-C score of 4.80. This stark contrast confirms that CTS provides the prerequisite foundational capability for SyncCFG to function effectively.

Necessity of Individual Modules. When evaluating the remaining modules, we observe that removing any single component from the full Harmony framework consistently degrades synchronization performance. Specifically, removing the Global-Local Decoupled Interaction module (*w/o GLDI*) drops the Sync-C score from 6.51 to 5.30. Similarly, removing the RoPE Alignment (*w/o RoPE Align*) decreases the score to 5.60, and removing SyncCFG (*w/o SyncCFG*) reduces it to 5.09. These results corroborate our core findings: CTS provides the robust alignment priors, GLDI and RoPE ensure precise architectural synchronization, and SyncCFG effectively amplifies these learned features during generation.

Table 4. **Leave-One-Out Ablation Study.** We evaluate the impact of removing individual components from the full Harmony model on the human-speech dataset. Note that applying SyncCFG without the foundational CTS training leads to severe artifacts and a performance collapse.

Model Configuration	Sync-C ↑
w/o SyncCFG	5.09
w/o RoPE Align	5.60
w/o GLDI	5.30
w/o CTS (w/o SyncCFG)	4.80
w/o CTS (w/ SyncCFG)	1.13
Full Model (Harmony)	**6.51**

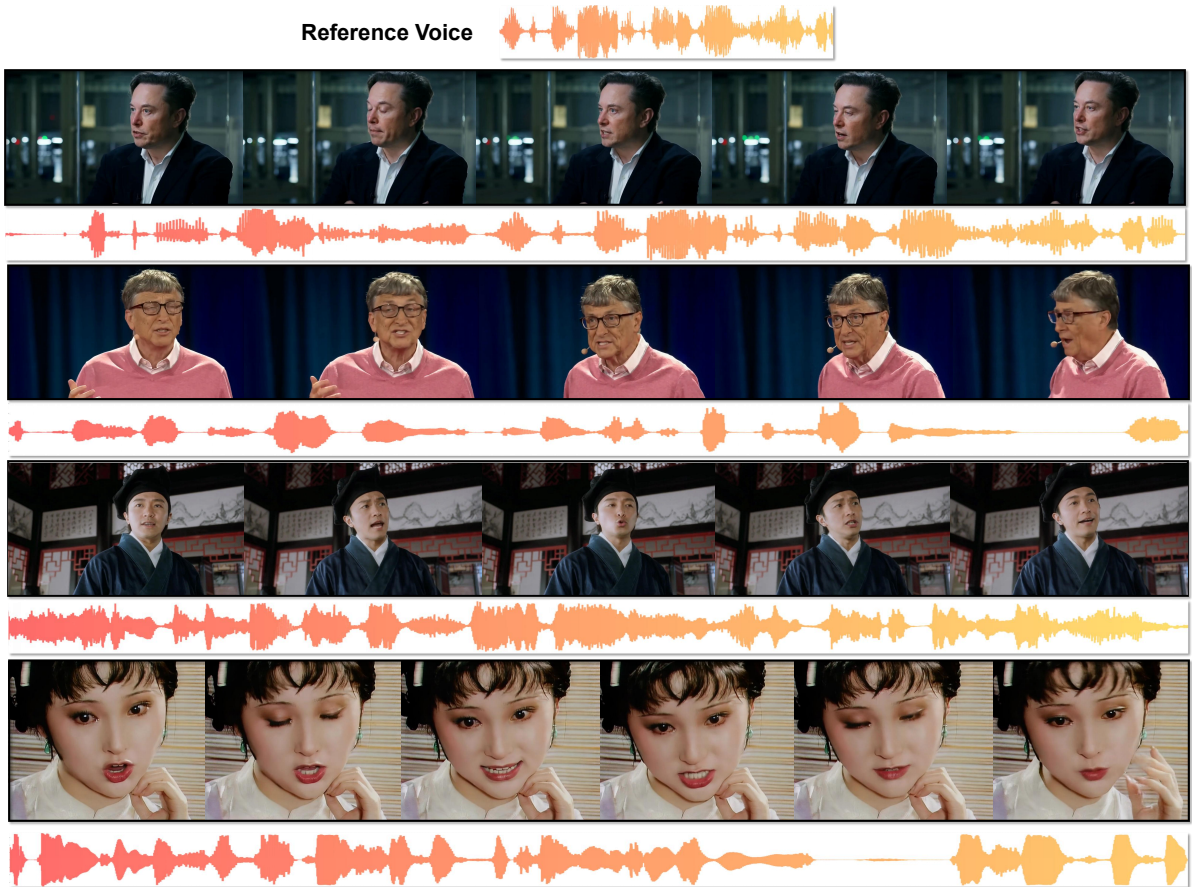

Figure 4. Visualization of the voice-clone results of our model.

8. Analysis of Off-Screen Audio and SyncCFG Behavior

In this section, we provide further insights into how Harmony processes complex audio-visual scenarios, specifically focusing on its behavior when handling off-screen sounds, such as background narration or unseen ambient noises. A common challenge in strongly aligned joint generation models is the risk of "forced alignment," where the network incorrectly binds off-screen audio to irrelevant on-screen visual elements (e.g., forcing a closed mouth to move when a voiceover is playing). We demonstrate that our architecture, equipped with RoPE-Aligned Attention and SyncCFG, naturally avoids this failure mode.

Audio-to-Video Cross-Attention. Because our proposed RoPE-Aligned Frame-wise Attention explicitly governs the cross-modal interaction, visualizing its attention maps provides direct insight into the model's spatial and temporal focus. As illustrated in Figure 5 (Left), when generating a scene with an *on-screen* sound source (e.g., a visible speaker), the attention map correctly converges on the corresponding visual region, such as the subject's mouth. Conversely, when the condition involves an *off-screen* narration, the attention map remains diffuse and uniformly distributed across the frame. It exhibits no specific spatial tendency, confirming that the model effectively recognizes the absence of a visible sound source and refrains from incorrect visual binding.

SyncCFG Guidance Heatmap. To further validate this robust behavior during inference, we analyze the spatial activation of our Synchronization-Enhanced CFG (SyncCFG). We visualize the SyncCFG guidance heatmap, formulated as the difference between the joint prediction and the driven negative anchor ($\hat{\epsilon}_\theta^{joint} - \hat{\epsilon}_\theta^{driven}$). High-value regions in this heatmap indicate areas where SyncCFG actively amplifies the alignment signal.

As shown in Figure 5 (Right), for off-screen audio inputs, the guidance heatmap does not highlight or amplify any incorrect local regions. Instead, it remains spatially agnostic, lacking any concentrated focus. This demonstrates that the "driven" negative anchors successfully function as intended: they provide a stable, static visual baseline that

Table 5. **Model Agnosticism Evaluation.** Comparison of the Harmony framework implemented with different video generation backbones. While the overall visual quality (AQ) is constrained by the base model's capacity, the synchronization performance (Sync-C) remains robustly high.

Model Configuration	AQ ↑	PQ ↑	Sync-C ↑
Harmony (Base: Wan2.1)	0.37	6.09	6.48
Harmony (Base: Wan2.2)	**0.48**	**6.20**	**6.51**

prevents SyncCFG from forcing spurious temporal alignments when the audio does not correlate with the visible on-screen dynamics.

9. Model Agnosticism

To demonstrate that the effectiveness of our proposed alignment mechanisms is not inherently tied to a specific network architecture, we evaluate the model agnosticism of the Harmony framework. In the main paper, the video generation branch of our model adapts the pre-trained Wan2.2 backbone. In this supplementary experiment, we replace this backbone with an earlier, less powerful version (Wan2.1) while keeping our proposed modules—specifically the Global-Local Decoupled Interaction (GLDI) module and Synchronization-Enhanced CFG (SyncCFG)—intact.

The quantitative results are presented in Table 5. As expected, utilizing a weaker base model results in a noticeable decrease in the generated visual Aesthetic Quality (AQ drops from 0.48 to 0.37) and a minor degradation in Audio Quality (PQ). These reductions are fundamentally limited by the generative capacity of the Wan2.1 backbone itself.

However, the crucial audio-visual synchronization performance remains highly consistent. The Sync-C score experiences only a negligible drop (from 6.51 to 6.48). This robust performance confirms that our proposed framework effectively captures and enforces cross-modal alignment independent of the base model's visual rendering capabilities.

10. Details about Voice Clone

In this section, we provide additional details on the voice cloning capability of our model, which is achieved through the use of a reference audio input, A_r. The mechanism begins by processing a short reference audio clip (typically 1-3 seconds) containing the desired voice timbre with our pre-trained audio VAE encoder [2]. This yields a compact latent representation, z_r, which effectively captures the unique, time-invariant characteristics of the speaker's voice while discarding the original phonetic content. As described in our main methodology, this reference latent z_r is then prepended to the noisy target audio latent $z_{a,t}$ during each step of the denoising process. By conditioning the MM-DiT on this fixed reference latent, the model is guided to synthesize new speech—based on the phonetic content from the transcript T_s—in the desired target voice.

To qualitatively validate the effectiveness of this approach, we provide examples in Figure 4. The figure demonstrates that our model can successfully clone a variety of distinct voice timbres onto newly generated speech content. Importantly, this high-fidelity voice cloning is achieved without degrading the visual quality of the generated video. The lip movements remain precisely synchronized with the cloned audio, and the overall facial expressions and video coherence are maintained at a high level. This highlights the model's ability to disentangle audio timbre from other generation aspects, enabling robust voice cloning within a coherent audio-visual output.

11. Audio-Driven Performance

As detailed in our main paper, our Cross-Task Synergy Training strategy is fundamental to the model's performance. A key component of this strategy is the inclusion of a deterministic, audio-driven video generation task, represented by the loss term $\mathcal{L}_{\text{driven}}^{\text{audio}}$. During training, this task explicitly requires the video branch to generate video conditioned on the clean, non-noisy audio latent $z_{a,0}$ (i.e., the audio latent at timestep $t_a = 0$). By directly optimizing for this objective, our model is inherently equipped with the ability to perform high-fidelity audio-driven video synthesis at inference time, making it a native capability rather than an emergent one.

To demonstrate the effectiveness of this native capability, we present qualitative results for audio-driven video generation in Figure 6. The figure showcases examples where video is generated solely from a target speech audio clip. The results exhibit high visual quality, characterized by natural facial expressions and coherent head movements. More importantly, the lip movements are precisely and accurately synchronized with the nuances of the input speech, validating the strong audio-visual alignment instilled by our training approach. This confirms that our Cross-Task Synergy strategy not only enhances joint generation but also directly enables high-fidelity, single-modality-driven applications.

12. More qualitative results

To further demonstrate the capabilities and robustness of our model, we present additional qualitative results organized into three key areas: generating high-quality human speech videos, rendering diverse artistic styles, and synthesizing complex ambient sounds.

More results on human speech. First, we showcase additional results on generating human speech videos in Figure 7. These examples highlight the model's ability to produce highly realistic talking heads with natural facial expressions and coherent movements. The synthesized speech

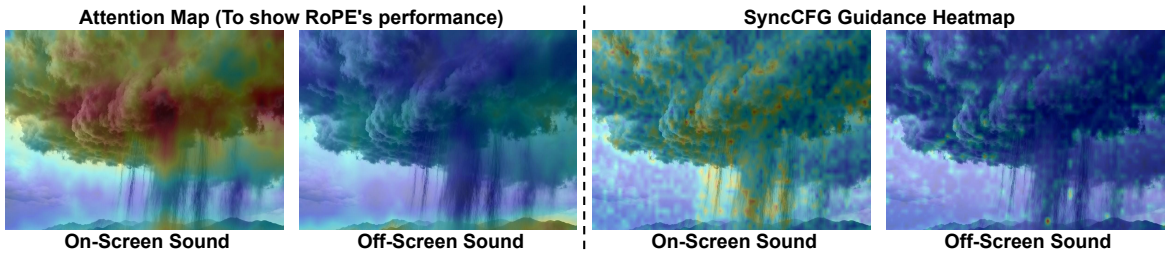

Figure 5. **Visualization of Model Behavior for On-Screen vs. Off-Screen Audio. Left:** Audio-to-Video cross-attention maps. The model highly focuses on the active sound source for on-screen audio but remains diffuse for off-screen narration. **Right:** SyncCFG guidance heatmaps ($\hat{\epsilon}_\theta^{joint} - \hat{\epsilon}_\theta^{driven}$). The guidance actively enhances specific regions for visible sound sources but appropriately shows no spatial concentration for off-screen sounds, avoiding forced alignment.

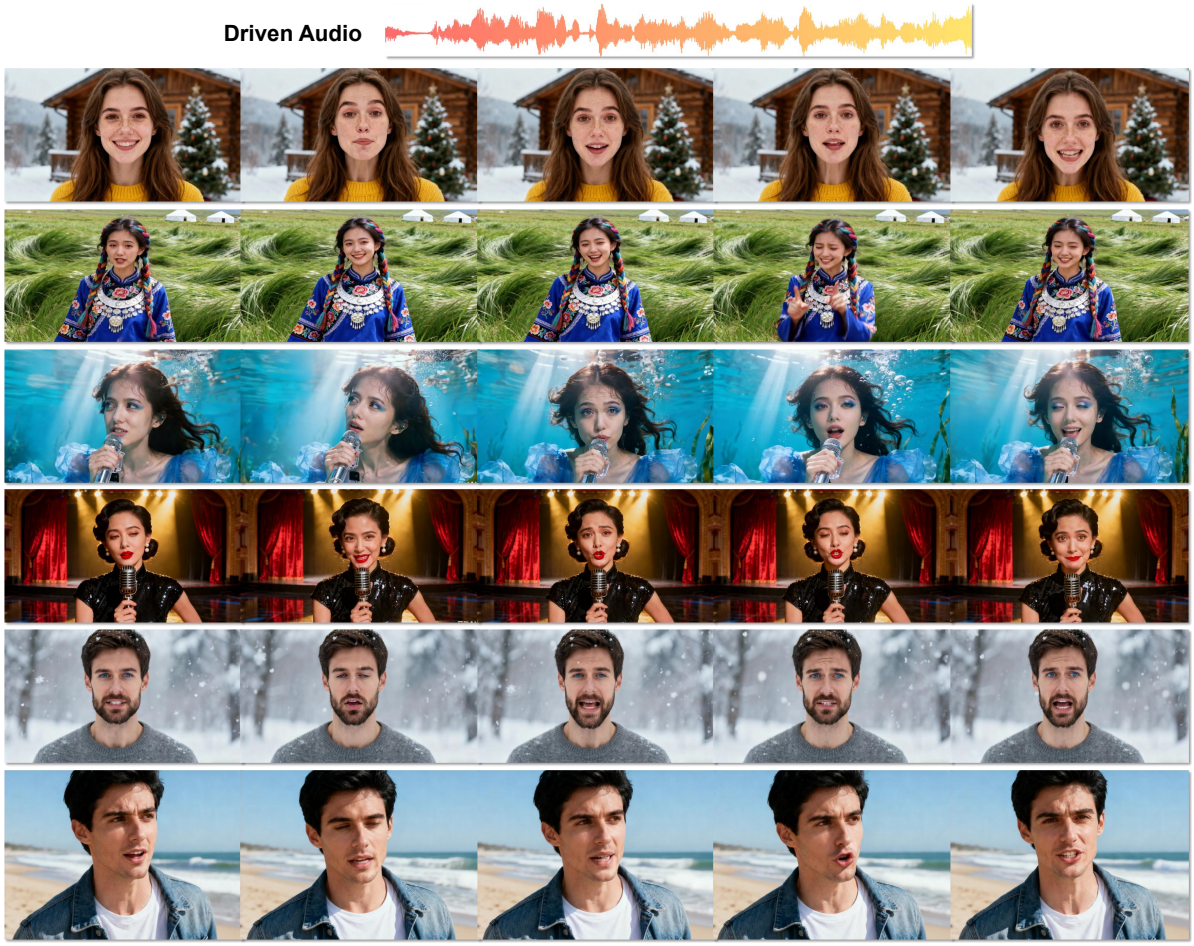

Figure 6. Visualization of the audio-driven results of our model.

is characterized by its clarity and natural prosody, capturing a range of vocal tones. Crucially, we maintain precise lip synchronization across all examples, which is fundamental for creating believable human speech. These results reinforce our model's core capability in generating high-quality, well-synchronized audio-visual speech content across various identities.

Diverse visual styles. Beyond photorealism, a key strength of our model is its capacity to generate video content across a wide spectrum of artistic styles. As illustrated in Figure 8, our model can produce outputs in distinct aesthetics such as Disney-style animation and traditional ink wash painting. These stylized generations maintain high visual quality, characterized by sharp details, vibrant colors, and tem-

porally coherent motion consistent with the target aesthetic. This demonstrates the model's flexibility in capturing and rendering complex artistic attributes.

Diverse Ambient Sounds. Our model demonstrates a remarkable capability to generate a wide spectrum of ambient sounds, extending beyond simple environmental noise. As illustrated in Figure 9, it can produce diverse and complex acoustic events—from the sharp, percussive bursts of fireworks to the structured harmonies of music. Crucially, each sound is rendered with high fidelity and meticulously synchronized with its corresponding visual source. This ability to construct rich, thematically consistent auditory environments validates our model's strength in enhancing the overall visual narrative.

Collectively, these examples validate our model's comprehensive generation capabilities. From producing highly synchronized human speech to rendering diverse artistic styles and creating rich, context-aware ambient soundscapes, our model demonstrates remarkable versatility. The ability to master these distinct yet complementary domains underscores its potential for creating highly expressive and immersive audio-visual content, pushing the boundaries beyond conventional generation methods.

References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1

[2] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911, 2025. 7

[3] 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. 3

[4] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 1

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 3

[6] Google. Gemini. https://gemini.google.com/, 2025. 1

[7] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE, 2024. 1

[8] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2

[9] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024. 3

[10] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 2

[11] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 1

[12] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7752–7762, 2025. 1

[13] Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025. 2, 3

[14] Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. 3

[15] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. 1

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 2

[17] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023. 3, 4

[18] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa,

Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. 2

[19] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2

[20] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025. 2

[21] Aesthetic Predictor V2.5. Aesthetic predictor v2.5. https://github.com/discus0434/aesthetic-predictor-v2-5, 2024. 2

[22] Ang Wang, Baole Ai, and et al Bin Wen. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1

[23] 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. 2, 3

[24] Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025. 1

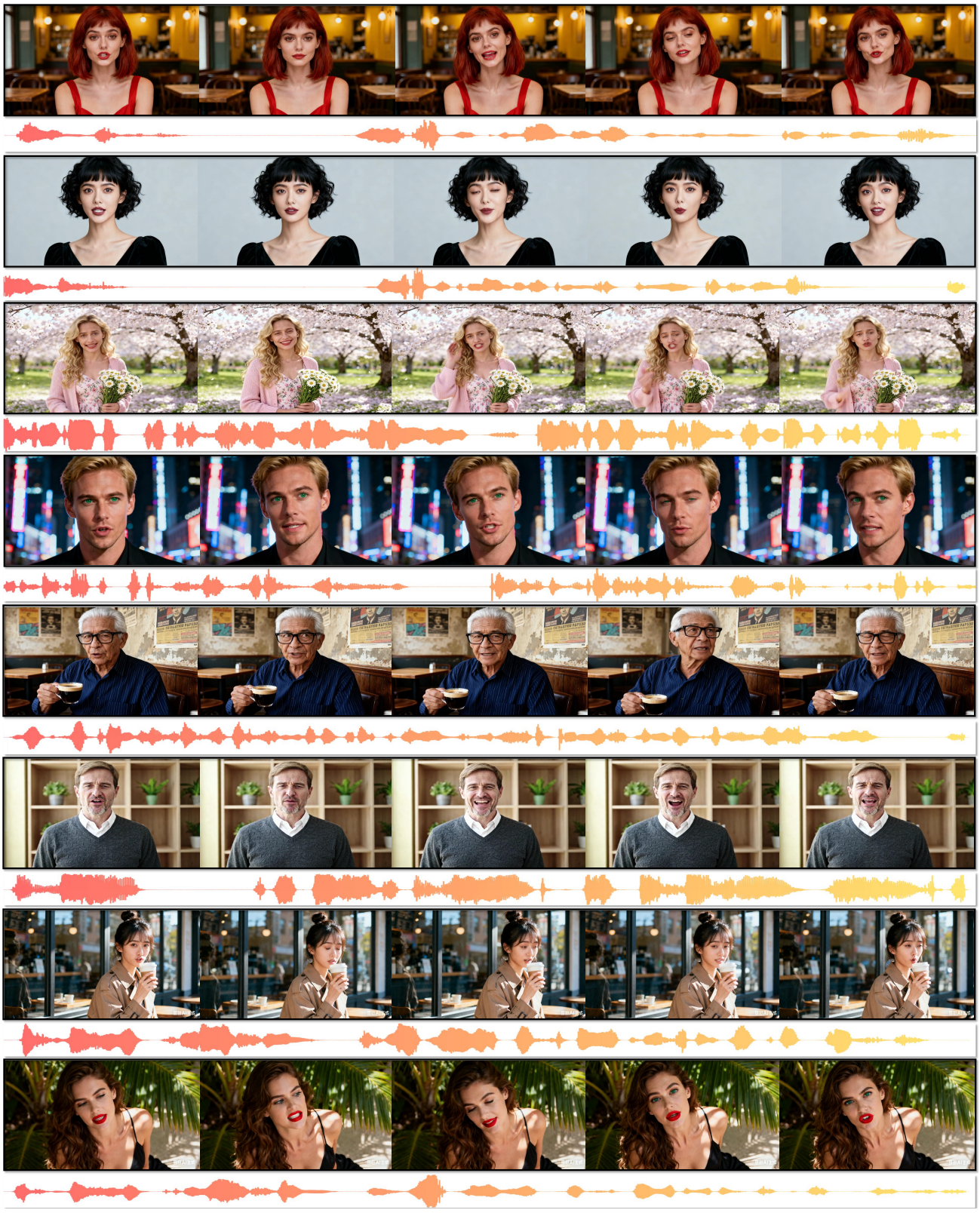

Figure 7. More results on human-speech video generation.

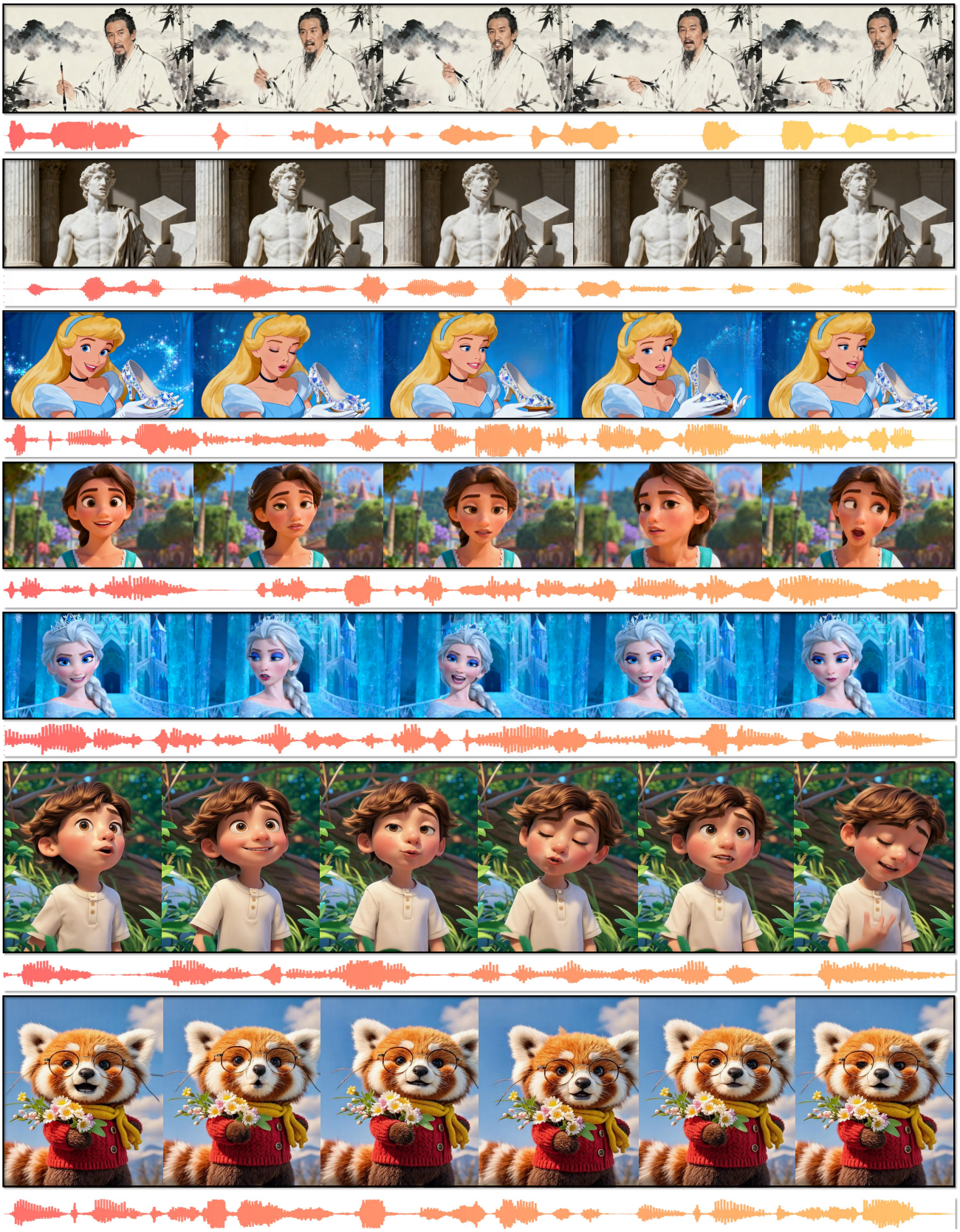

Figure 8. Visualization of speech-video generation in diverse style.

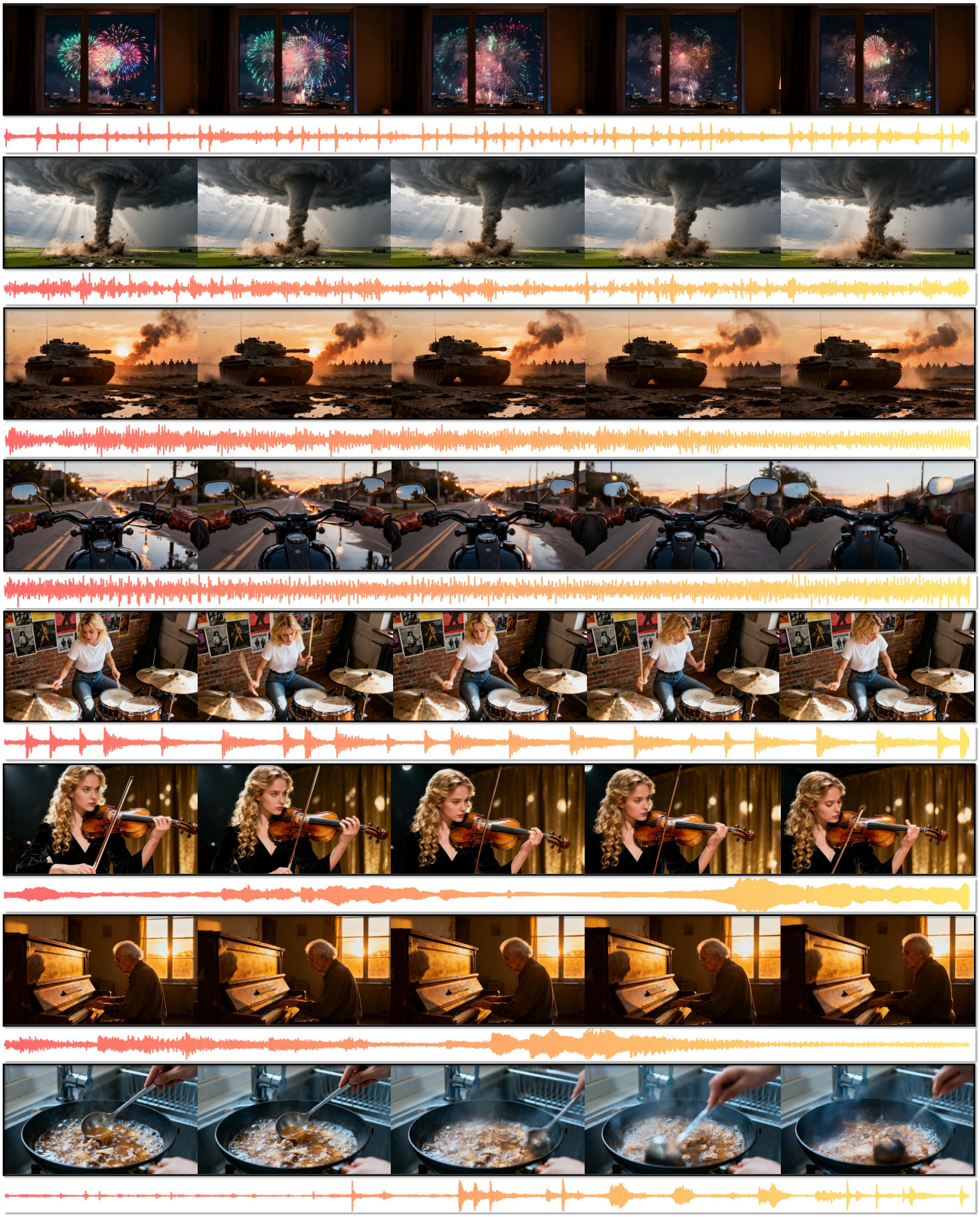

Figure 9. More results on ambient-sound video generation.