TARO: Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning for Synchronized Video-to-Audio Synthesis

Supplementary Material

OAC	TRA	FD↓	FAD↓	FID↓	IS↑	KL↓	Acc(%) ↑	CLIP ↑
		0.55	1.24	8.80	50.01	5.97	97.04	13.73
	✓	0.51	0.78	8.65	52.43	5.86	97.13	13.58
✓		0.53	1.10	8.85	52.65	5.98	97.16	14.05
✓	✓	0.47	0.94	8.21	56.60	5.71	97.19	14.10

Table 7. **Ablation study on key components of TARO.**

	FD↓	FAD↓	FID↓	IS↑	KL↓	Acc(%) ↑	CLIP ↑
w/o weighted	0.51	0.97	8.34	53.36	5.84	97.15	14.04
w/ weighted **(Ours)**	0.47	0.94	8.21	56.60	5.71	97.19	14.10
Global Pooling	0.51	1.16	9.21	50.53	5.94	97.06	14.01
Interpolate	0.49	1.09	8.23	57.01	5.73	97.13	14.12
Conv **(Ours)**	0.47	0.94	8.21	56.60	5.71	97.19	14.10

Table 8. **Ablation study on the details of Timestep-Adaptive Representation Alignment.**

Encoder	FD↓	FAD↓	FID↓	IS↑	KL↓	Acc(%) ↑	CLIP ↑
wav2vec 2.0 [2]	0.48	1.41	8.90	56.36	5.88	97.15	14.04
CLAP [38]	0.49	0.81	8.44	52.02	5.53	96.58	13.85
BEATs [5]	0.52	0.74	8.25	53.31	5.64	97.13	13.41
EAT [6] **(Ours)**	0.47	0.94	8.21	56.60	5.71	97.19	14.10

Table 9. **Ablation study on different audio encoders.**

Depth	FD↓	FAD↓	FID↓	IS↑	KL↓	Acc(%) ↑	CLIP ↑
2	0.49	1.49	8.52	56.57	5.73	97.13	14.14
4 **(Ours)**	0.47	0.94	8.21	56.60	5.71	97.19	14.10
6	0.48	0.82	8.71	53.02	5.83	97.11	13.75
8	0.48	0.77	8.51	54.39	5.72	96.92	13.73
10	0.49	0.79	8.59	54.26	5.79	97.09	13.63

Table 10. **Ablation study on the effect of injecting audio encoder features at different transformer blocks.**

A. Demo Audios

We recommend that readers refer to our project page at github.com/triton99/TARO, showcasing extensive qualitative comparisons between our TARO and SOTA video-to-audio generation methods [15, 24, 29, 32, 36, 39, 45]. Please note that since the provided project page for this supplementary material is *offline*, and therefore, *no modifications can be made after submission*; it is offered solely for the convenience of visualization. The project page features various demo video-to-audio synthesis, including comparisons for our TARO and prior methods [15, 24, 29, 32, 36, 39, 45] on both VGGSound [3] and Landscape [18] datasets.

B. Subjective evaluation

To comprehensively evaluate the performance of our Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning (TARO) in video-to-audio synthesis, we conduct a user study involving 20 videos from the VGGSound dataset [3] and 10 videos from the Landscape dataset [18]. Our study focuses on two key aspects: audio quality and temporal alignment. For audio quality assessment (Mean Opinion Score for Audio Quality, MOS-Q), participants are instructed to focus exclusively on the generated audio, disregarding the accompanying video content, and rate its perceptual quality. This ensures that judgments reflect the naturalness and clarity of the audio without being influenced by visual cues. For temporal alignment evaluation (Mean Opinion Score for Content Alignment, MOS-A), participants assess how well the generated audio synchronizes with the corresponding video, ignoring audio quality to provide an unbiased measure of synchronization accuracy. To mitigate potential biases, we group samples by video and randomly shuffle their order within each group before presenting them to participants. Each sample is rated on a 1–5 Likert scale [19], offering a robust subjective evaluation of TARO's capability to generate high-quality and well-aligned audio in video-to-audio synthesis. Our study involved 20 participants, both male and female, aged from 24 to 30, primarily graduate students, with 50% having experience in generative models.

C. Comprehensive Metric Analysis for Ablation Study

In addition to the primary metrics discussed in the main paper, Table 7 presents a detailed comparison of FID, IS, and KL scores across different ablation settings. The results show that Timestep-Adaptive Representation Alignment (TRA) improves distributional matching, reducing FID and KL while increasing IS, indicating enhanced generative quality. Onset-Aware Conditioning (OAC) primarily benefits perceptual alignment, contributing to a higher IS score. The full TARO, integrating both components, achieves the best performance across all metrics, highlighting their complementary role in improving fidelity, synchronization, and generative quality in video-to-audio synthesis.

Ablation on Timestep-Adaptive Representation Alignment As shown in Table 8, FID improves with adaptive weighting and convolution-based projection, achieving the lowest value and indicating better distribution alignment with real audio. IS is maximized when interpolation is used, suggesting enhanced diversity, but it comes at the cost of weaker distribution matching. Our convolution-based projection achieves a strong balance, maintaining a high IS score while minimizing KL divergence, demonstrating improved synthesis quality and robustness in audio generation.

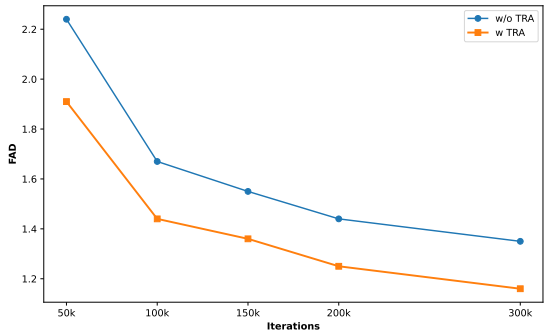

Figure 7. **Training efficiency.**

Depth	FD↓	FAD↓	Acc(%) ↑	CLIP ↑
Energy control	0.56	0.89	96.10	13.92
Down-sampled mel	0.59	1.02	96.90	13.88
Onset cues (Ours)	0.47	0.94	97.19	14.10

Table 11. **Ablation study on temporal features.**

Ablation on Different Audio Feature Extractor Table 9 shows that EAT [6] achieves the lowest FID, indicating superior distribution alignment with real audio compared to other encoders. IS scores vary across encoders, with EAT [6] and wav2vec 2.0 [2] performing best, suggesting they enable more diverse and expressive audio generation. KL divergence is lowest with CLAP [38], which may indicate smoother latent representations, though its lower accuracy and CLIP Score suggest weaker synchronization. EAT [6] maintains a strong balance, achieving competitive KL while excelling in FID and IS, reinforcing its effectiveness in improving fidelity and synchronization.

Ablation on Audio Encoder Injection Depth Table 10 demonstrates that injecting at the 4th block yields the lowest FID, confirming its effectiveness in aligning the learned distribution with real audio. While IS remains relatively stable, early injection (2nd block) attains the highest score but also results in a higher FID, indicating potential misalignment. KL divergence is lowest at the 4th block, suggesting it provides the best trade-off between audio-visual integration and representation refinement. Later injections (6th, 8th, and 10th blocks) slightly increase FID and KL, implying reduced effectiveness in synchronizing latent audio features with the video context.

D. Additional Results & Analysis

Training efficiency As shown in Fig. 7, TRA improves both training stability and generation quality compared to the vanilla model.

Comparison with other temporal features. Tab. 11 shows the effectiveness of different temporal features: energy control [16], down-sampled melspectrogram [35], and onset cues (Ours). We train all methods with 500k iterations. As shown, onset cues achieve the best balance, yielding the highest synchronization accuracy and CLIP score, with competitive FAD and FD.

Method	IB↑	Onset Acc↑	Onset Acc↑	Offsets(s)↓
SpecVQGAN	56.43	28.09	54.70	1.18
Im2Wav	55.75	27.76	52.57	1.16
Diff-Foley	59.64	25.47	61.90	1.09
See & Hear	62.76	26.41	60.72	1.20
FoleyCrafter	63.66	29.18	55.19	1.14
Frieren	61.26	26.53	63.72	0.98
MDSGen	60.78	27.90	58.14	1.16
Ours	64.22	29.37	62.17	0.97

Table 12. **Other metrics on VGGSound dataset.**

Other metric report. In addition to commonly reported metrics, we provide some recently used metrics such as ImageBind Score, Onset Acc, Onset AP, and Temporal Offset in Tab. 12 on the VGGSound dataset.

E. Limitation

Despite the strong performance of TARO, several limitations remain. First, while the VGGSound [3] dataset provides a diverse set of audio-visual samples, its scale may not fully exploit the potential of our approach. Expanding to larger and more varied datasets could enhance generalization. Second, our method is currently constrained to a fixed video length, limiting its adaptability to variable-length sequences or practical applications. Third, like existing video-to-audio synthesis models, TARO is primarily designed to generate Foley sounds and struggles with handling human speech, requiring more fine-grained linguistic and phonetic understanding. Additionally, the effectiveness of the Onset-Aware Conditioning (OAC) module depends on the accuracy of onset detection models. Errors in onset prediction could lead to misalignment in synthesized audio. Addressing these limitations presents exciting directions for future research, including improving adaptability and generalization and expanding beyond Foley sound generation.

F. Additional Visualizations

Figs. 8, 9, and 10 present spectrogram visualizations compare our method with prior works [15, 24, 29, 32, 36, 39, 45]. These visualizations illustrate how different models capture event-driven acoustic cues and maintain synchronization with visual content. Our method demonstrates more precise temporal alignment, clearer event transitions, and improved spectral fidelity compared to baselines, which often exhibit artifacts, misaligned sound events, or missing key audio cues. These results further validate the effectiveness of our proposed framework in generating high-quality, temporally coherent audio.

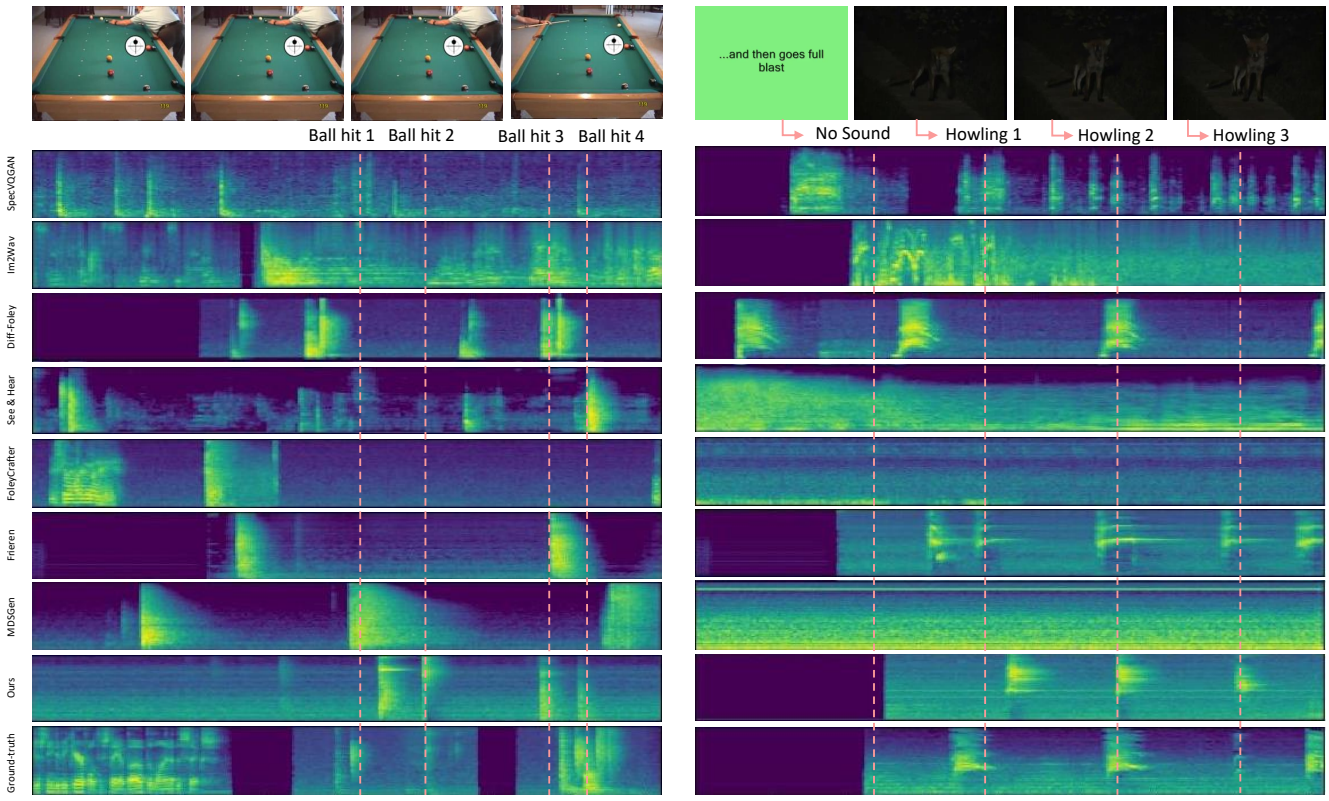

Figure 8. **Visual comparisons on *1mMqLP36sCQ_000245* and *1JsIcP2nXMw_000108* from the VGGSound dataset.**

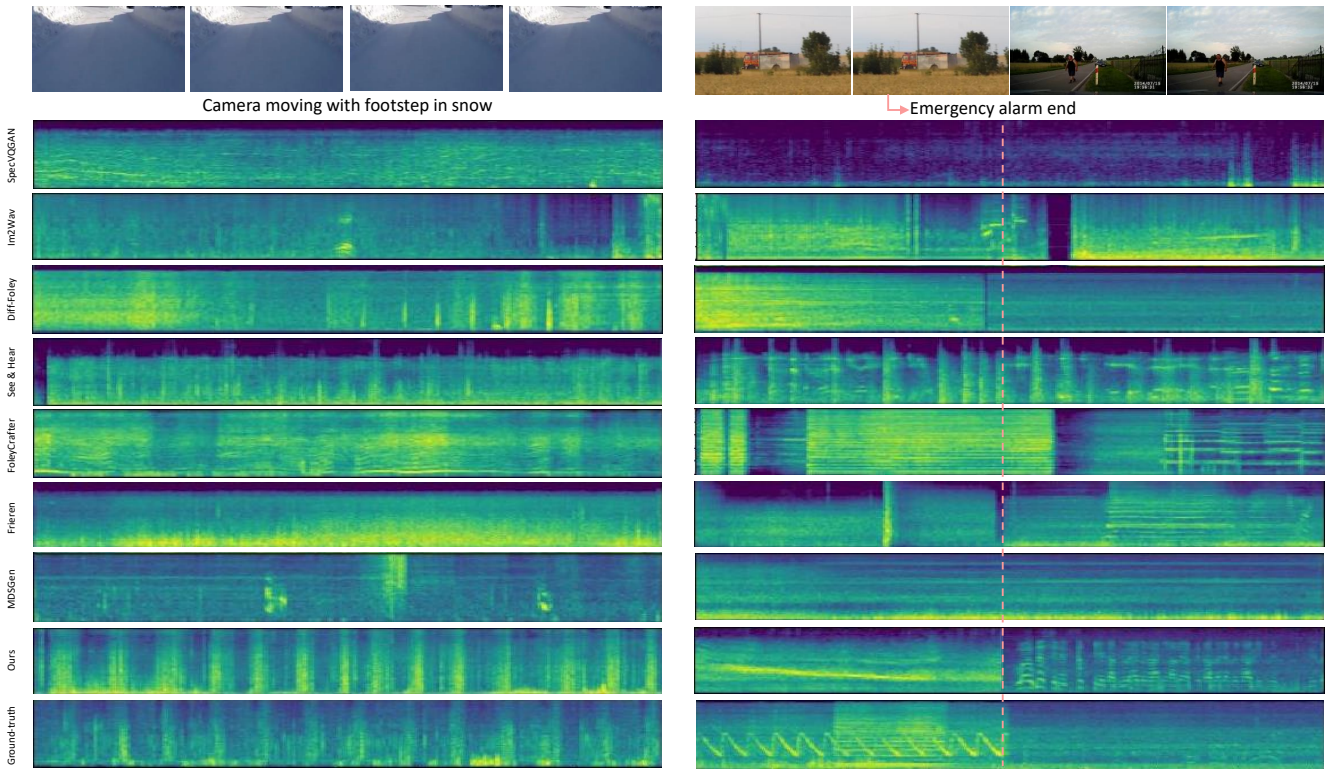

Figure 9. **Visual comparisons on *KydSULgAHFI_000084* and *0N6S5OoG7Vg_000150* from the VGGSound dataset.**

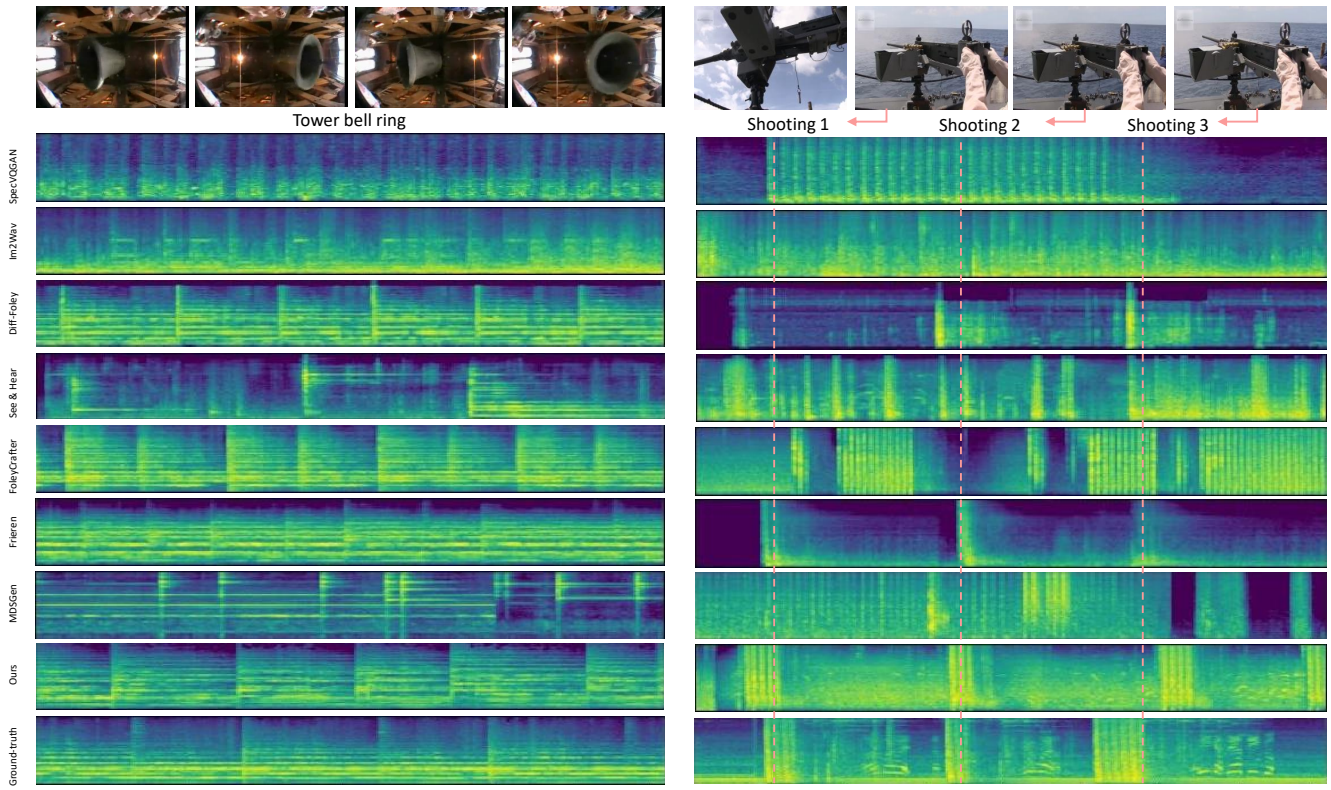

Figure 10. **Visual comparisons on *6vl7eSBL-ag_000090* and *OFOHNgpDS38_000091* from the VGGSound dataset.**