



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

Tri Ton Ji Woo Hong Chang D. Yoo

Korea Advanced Institute of Science and Technology (KAIST)

{trith, jiwoohong93, cd_yoo}@kaist.ac.kr

Abstract

This paper introduces *Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning (TARO)*, a novel framework for high-fidelity and temporally coherent video-to-audio synthesis. Built upon flow-based transformers, which offer stable training and continuous transformations for enhanced synchronization and audio quality, TARO introduces two key innovations: (1) *Timestep-Adaptive Representation Alignment (TRA)*, which dynamically aligns latent representations by adjusting alignment strength based on the noise schedule, ensuring smooth evolution and improved fidelity, and (2) *Onset-Aware Conditioning (OAC)*, which integrates onset cues that serve as sharp event-driven markers of audio-relevant visual moments to enhance synchronization with dynamic visual events. Extensive experiments on the VGGSound and Landscape datasets demonstrate that TARO outperforms prior methods, achieving relatively 53% lower Frechet Distance (FD), 29% lower Frechet Audio Distance (FAD), and a 97.19% Alignment Accuracy, highlighting its superior audio quality and synchronization precision. The code is available at: github.com/triton99/TARO.

1. Introduction

Video-to-audio generation has garnered increasing attention due to its broad applications in content creation and film production [1, 7]. The goal is to synthesize audio that is both high-quality and temporally aligned with video frames. However, achieving high-fidelity and precisely synchronized audio remains challenging, especially when dealing with rapidly changing visual content.

Despite progress in video-to-audio generation, existing approaches struggle to balance fidelity, synchronization, and efficiency. GAN-based methods [4] improve audio fidelity but often fail to maintain temporal coherence. Transformer-based autoregressive methods [13, 27] offer greater modeling expressiveness but suffer from slow in-

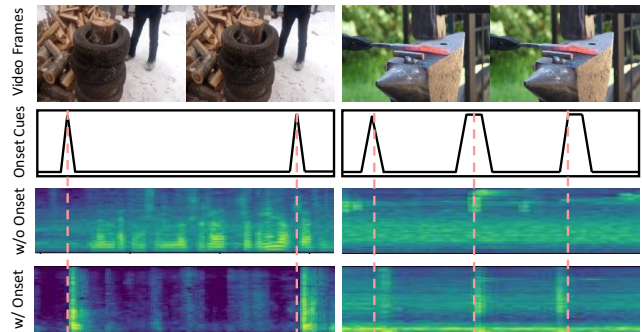


Figure 1. Our TARO leverages onset-aware conditioning to improve synchronization, aligning generated audio with event-driven cues in video.

ference and accumulating timing errors. Diffusion-based methods [22, 25, 34, 40] improve fidelity through latent diffusion and cross-modal pretraining, yet their high computational costs hinder real-time deployment. In contrast, flow-based methods [9, 35] provide stable training and continuous transformations, emerging as a promising alternative for efficient video-to-audio synthesis.

Existing video-to-audio approaches naively learn latent audio representation through reconstruction loss, which can lead to suboptimal quality. Recently, REPA [39] has attempted to leverage pretrained features [24] as a constraint to guide latent representations in image synthesis. Inspired by this, we explore applying a similar constraint-driven approach to latent audio representation learning. Directly applying methods like REPA is non-trivial, and we carefully design TRA by (1) dynamically adjusting the alignment strength based on the diffusion noise schedule, (2) addressing sequence length mismatch, and (3) selecting appropriate audio encoder outputs. This prevents over-regularization while preserving semantic grounding throughout training. Yet adapting this approach for video-to-audio synthesis is challenging, as imposing strict constraints on latent representations in early timesteps, when noise dominates, can hinder learning rather than improve alignment. Instead, a more adaptive strategy is required to gradually refine align-

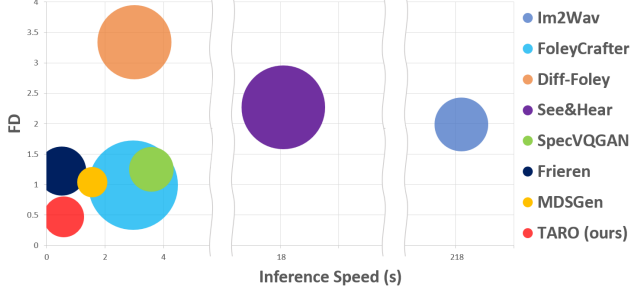


Figure 2. Comparison of FD (y-axis) and inference speed (x-axis) across different models, with marker size representing parameter size.

ment as the latent representation evolves. Additionally, achieving precise synchronization at the event level remains an open challenge. Existing methods struggle to capture fine-grained auditory cues tied to specific visual events, often generating misaligned or unnatural sounds. This issue arises from the lack of explicit mechanisms to detect and condition on key auditory events such as footsteps, object interactions, or percussive impacts, which serve as strong temporal anchors for realistic audio synthesis.

To address these challenges, we introduce Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning (TARO), a novel flow-based transformer framework that enhances both fidelity and synchronization in video-to-audio generation. TARO introduces two key innovations to improve latent audio representation and event-level alignment. First, Timestep-Adaptive Representation Alignment (TRA) dynamically adjusts alignment strength based on the noise schedule, ensuring progressive refinement of latent representations during training. Second, Onset-Aware Conditioning (OAC) explicitly integrates event-driven cues to align generated audio with video motion by detecting key auditory events. These sparse and semantically meaningful onset cues serve as temporal anchors, guiding the model to generate sound at appropriate moments. This conditioning mechanism enhances temporal precision, improving both perceptual quality and the naturalness of video-to-audio synthesis. Extensive experiments on two public datasets, VGGSound [3], and Landscape [16] datasets, demonstrate that TARO synthesizes high-quality, well-aligned audio, outperforms state-of-the-art (SOTA) methods across various comprehensive metrics as shown in Figure 2.

In summary, our contributions are as follows:

- We propose **Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning (TARO)**, a novel flow-based transformer framework for video-to-audio generation. TARO integrates both visual and onset-based control signals to enhance the fidelity and synchronization of generated audio.
- We introduce a novel **Timestep-Adaptive Representation**

Alignment strategy that dynamically adjusts alignment strength based on the noise schedule, refining latent representations for high-quality, temporally coherent audio.

- We present an **Onset-Aware Conditioning** that captures key event-driven cues to refine synchronization, enabling precise audio alignment with dynamic visual events.

2. Related Works

Video-to-audio generation. Video-to-audio generation aims to synthesize temporally synchronized and semantically relevant audio from video. Transformer-based autoregressive methods like SpecVQGAN [13] offer flexibility but suffer from slow inference and weak alignment. Im2Wav [27] extends this to images via CLIP [26] embeddings but struggles with dynamic video content. Diffusion-based models [12, 28, 36–38] improve quality but require multiple steps and large parameters. Diff-Foley [22] enhances alignment via latent diffusion, while See & Hear [34] improves multimodal integration but lacks fine-grained temporal precision. FoleyCrafter [40] incorporates a temporal controller for high-quality Foley sound but remains computationally heavy. Frieren [31] reduces steps via rectified flow matching but may miss sparse, event-driven dynamics. MDSGen [25] employs masked diffusion for efficiency but still relies on multiple steps for fidelity. Furthermore, several works incorporate auxiliary temporal features as condition information to guide the generation process. TiVA [30] leverages smoothed energy control, and ReWaS [14] uses down-sampled mel-spectrograms. However, these signals often introduce temporal redundancy or smoothing artifacts, making them less effective at capturing fine-grained event boundaries critical for precise alignment. Our TARO framework addresses these issues using timestep-adaptive alignment, and onset-aware conditioning to enhance audio quality, synchronization, and computational efficiency.

Flow-based generative models. Flow-based generative models have gained traction for their efficiency and stability in image and audio synthesis, providing an alternative to score-based diffusion models. Flow matching [18] models a continuous vector field to smoothly transport samples from noise to data using ordinary differential equations (ODEs), achieving robust training and superior performance. In audio synthesis, Voicebox [15] and Audiobox [29] leverage flow matching for speech and unified audio generation, while VoiceFlow [10] applies rectified flow matching for efficient text-to-speech. Our work bridges this gap by extending flow-based transformers to video-to-audio synthesis to improve generation quality and synchronization.

Audio Representation Learning. Recent advances in audio representation learning have significantly improved the ability to extract meaningful features for audio understanding, offering valuable insights for generative models. Self-supervised learning frameworks such as wav2vec 2.0 [2]

capture robust speech representations, while CLAP [33] employs contrastive language-audio pretraining to learn general-purpose audio semantics. Transformer-based models like BEATs [5] leverage masked spectrogram learning for audio classification, whereas EAT [6] introduces an efficient audio transformer trained with utterance-frame objectives. Despite their effectiveness, these models have yet to be fully leveraged for video-to-audio tasks, where precise temporal alignment and semantic consistency are critical. Our approach integrates pretrained audio encoders into a timestep-adaptive alignment mechanism, ensuring that latent representations evolve meaningfully across synthesis.

3. Preliminary

Flow-based generative models provide an alternative to Denoising Diffusion Probabilistic Models (DDPM) [12] by leveraging a continuous-time framework for data generation. Unlike DDPM, which employs discrete noise schedules, flow-based models [18, 23] define a smooth perturbation process over $t \in [0, 1]$, continuously evolving data from a noise distribution to a structured target distribution. To model this transformation, given data $x_* \sim p(x|c)$, conditioned on c , a perturbation process is defined as:

$$x_t = \alpha_t x_* + \sigma_t \epsilon, \quad (1)$$

where α_t is a decreasing function of t for controlling the data component and σ_t is an increasing function of t for governing the noise contribution, and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. This process follows a probability flow described by an ordinary differential equation (ODE):

$$\dot{x}_t = v(x_t, t, c), \quad (2)$$

where $v(x_t, t, c)$ represents the transport vector field guiding the flow from noise to data, conditioned on c .

To enable efficient sample generation, this ODE is numerically solved, such as using the Euler method, starting from random noise $\epsilon \sim \mathcal{N}(0, I)$. The transport vector field is parameterized as $v_\theta(x_t, t, c)$, typically implemented as a neural network, and trained using the conditional flow matching (CFM) objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, (x_*, c), \epsilon} [\|v_\theta(x_t, t, c) - (\dot{\alpha}_t x_* + \dot{\sigma}_t \epsilon)\|_2^2], \quad (3)$$

where $\dot{\alpha}_t x_* + \dot{\sigma}_t \epsilon$ represents the target velocity.

Our approach follows the stochastic interpolant framework proposed by [23], which unifies flow-based and diffusion models under a flexible interpolant design. We adopt a linear interpolant, defined as $\alpha_t = 1 - t$ and $\sigma_t = t$, which provides a simple yet effective mapping from noise to data over $t \in [0, 1]$. This choice ensures stable training while preserving temporal coherence in audio generation. Additionally, our approach is adaptable to other interpolants,

such as variance-preserving (VP) interpolants, defined by $\alpha_t = \cos(\frac{\pi}{2}t)$ and $\sigma_t = \sin(\frac{\pi}{2}t)$, which offer smoother noise schedules.

4. Method

4.1. Overview of TARO

Our approach builds upon the MMDiT [9], adapting it for video-to-audio synthesis by incorporating event-driven synchronization and representation alignment mechanisms. As illustrated in Figure 3, the pipeline begins with processing the input video through a pretrained visual encoder [22] with frozen parameters to extract frame-level features, denoted as $c_v \in \mathbb{R}^{L_v \times D_v}$ where L_v and D_v represent the number of frames and the visual feature dimension, respectively. These features capture the overall scene structure and serve as the primary conditioning input for audio synthesis. To enhance audio-visual synchronization, we introduce an Onset-Aware Conditioning (OAC), which employs a pretrained onset detection model [8] to generate onset cues $c_o \in \mathbb{R}^{1 \times L_o}$, identifying key moments where audio-relevant events occur in the video frames.

During training, a mel-spectrogram $M \in \mathbb{R}^{F \times T}$, where F represents the frequency bins and T denotes the temporal resolution, is encoded through a Variational Autoencoder (VAE) to produce a latent representation $z \in \mathbb{R}^{8 \times F/r \times T/r}$, where r is the downsampling factor. This latent representation is further patched and tokenized with a patch size of $p = 2$, yielding $z' \in \mathbb{R}^{L_z \times D_z}$, where L_z represents the sequence length of the tokens, and D_z is the hidden dimension. These tokenized latents are then processed by a flow-based multimodal transformer, which consists of N stacked multimodal transformer blocks. Throughout this process, the model integrates cross-modal attention by conditioning on visual features c_v and onset cues c_o , ensuring that the generated audio aligns with the video’s structural and event-driven information.

To enable effective multimodal integration, our framework concatenates the projected visual features and audio latent embeddings along the sequence dimension before feeding them into the transformer blocks, as illustrated in Figure 4. Given the differences in feature spaces and numerical scales between these modalities, Adaptive Layer Normalization (AdaLN) is applied separately to each modality, ensuring consistent processing and preventing imbalance between visual and audio representations. These processed tokens then interact through a joint attention mechanism, where query, key, and value representations from both modalities are concatenated and jointly attended, allowing the model to capture fine-grained audio-visual correspondences and refine the latent representation synchronized. A key component of our framework is the Timestep-Adaptive Representation Alignment (TRA), which dynam-

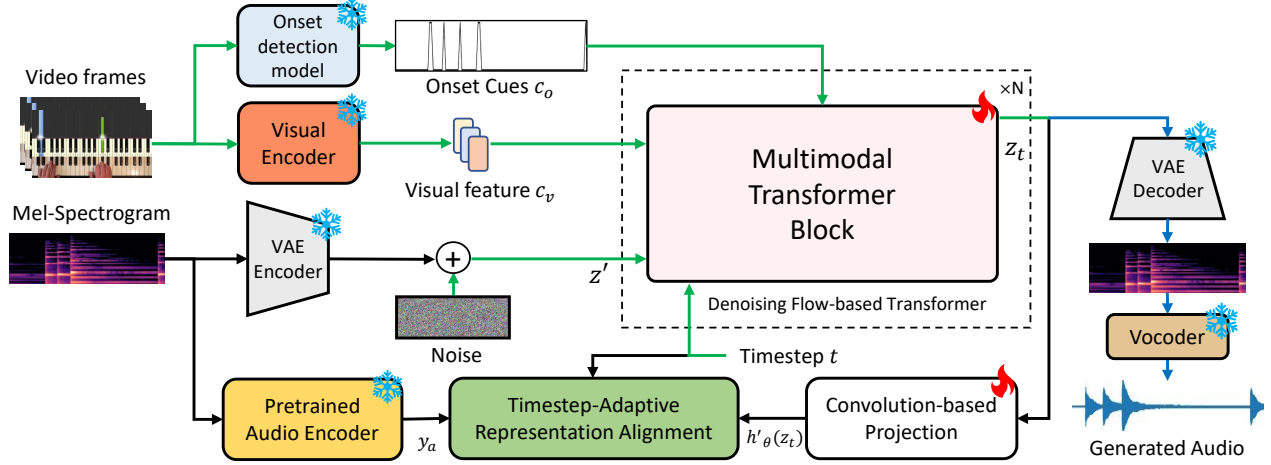


Figure 3. **Overview of TARO.** Our TARO is a flow-based multimodal transformer for video-to-audio generation, integrating Timestep-Adaptive Representation Alignment (TRA) and Onset-Aware Conditioning (OAC) to enhance synchronization and fidelity. Black arrows \rightarrow denote branches used only in training, blue arrows \rightarrow for inference only, and green arrows \rightarrow for both training and inference.

ically aligns latent representations with high-quality pre-trained audio encoders [2, 5, 6, 33]. This adaptive strategy leverages pretrained audio priors to enhance both audio fidelity and synchronization with visual content.

During inference, the model initializes the latent audio representation from Gaussian noise and iteratively refines it through the flow-based denoising process, guided by visual features c_v and onset cues c_o . The final latent is decoded by the VAE decoder to reconstruct the mel-spectrogram $\hat{M} \in \mathbb{R}^{F \times T}$, which is then converted into an audio waveform using a vocoder. The following sections detail the OAC in Sec. 4.2, followed by the TRA in Sec. 4.3.

4.2. Onset-Aware Conditioning

Accurate audio-video synchronization requires identifying event-driven cues that indicate key auditory moments. Onset cues capture abrupt sound changes, such as percussive strikes, footsteps, or sudden object interactions, providing structured timing signals crucial for aligning acoustic events with visual motion. While pretrained visual features [22] encode semantic content and broad temporal correlations, they lack explicit event timing. In contrast, onset cues introduce discrete temporal markers, refining fine-grained synchronization in video-to-audio generation. To incorporate these cues, we utilize a pretrained video onset detection model [8] to extract frame-wise onset events c_o , which highlight key transitions in the video. As illustrated in Figure 4, these cues undergo a projection module using multi-layer perceptions (MLPs) to ensure alignment with the latent space. To reinforce temporal coherence, the projected onset features are summed with timestep embeddings before integration into the network. This step allows the model to jointly encode both event timing and global progression through the generative process, ensuring more structured conditioning. We employ AdaLN to modu-

late both visual features and audio latents, serving as guidance signals in a lightweight and temporally aligned manner, enabling fine-grained adaptation across different video frames. AdaLN dynamically scales and shifts feature activations based on the onset-conditioned parameters, ensuring flexible processing throughout the generative process. This design avoids the computational overhead of token-level concatenation or cross-attention, which is commonly used in DiT-style methods. Specifically, given an input representation x and an onset-based conditioning signal c , AdaLN is computed as:

$$\text{AdaLN}(x|\gamma_c, \beta_c) = \gamma_c x + \beta_c, \quad (4)$$

where modulation parameters are obtained by MLP:

$$\gamma_c = \text{MLP}_\gamma(c), \quad \beta_c = \text{MLP}_\beta(c). \quad (5)$$

By integrating onset cues with timestep embeddings, the model captures both fine-grained event timing and the overall generative progression, ensuring stable synchronization throughout the synthesis process.

4.3. Time-Adaptive Representation Alignment

The standard alignment method [39] apply fixed constraints throughout synthesis, which can hinder learning in early timesteps when noise dominates. TRA dynamically adjusts alignment strength based on the noise schedule, allowing latent representations to evolve progressively. This adaptive approach prevents overly strict constraints in noisy stages while gradually reinforcing structure as synthesis progresses, ensuring smooth and effective alignment. Given a pretrained encoder f_{enc} and a clean audio sample x_a , we compute its embedding as $y_a = f_{enc}(x_a) \in \mathbb{R}^{L_a \times D_a}$, where L_a represents the number of patches and D_a denotes the embedding dimension. To align with this reference,

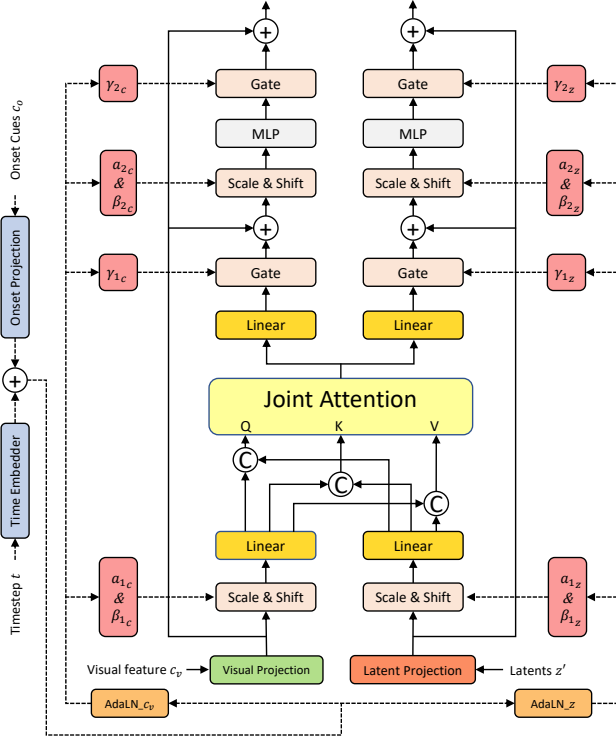


Figure 4. **One Multimodal Transformer Blocks.** The block integrates onset cues, visual features, and latent audio representations through adaptive modulation and joint attention. + denotes summation, and C represents concatenation.

we use a trainable projection head h_ϕ , which transforms the model latent representation $z_t \in \mathbb{R}^{L_z \times D_z}$ at timestep t into the same y_a feature space as $h_\phi(z_t) \in \mathbb{R}^{L_z \times D_a}$. This projection ensures that the learned representation matches the structured features of clean audio, enabling effective alignment. To achieve this alignment, we introduce a time-weighted alignment loss, adjusting alignment strength dynamically based on the noise schedule:

$$\mathcal{L}_{\text{align}} = \sum_t w(t) \cdot d(h'_\phi(z_t), y_a), \quad (6)$$

where $d(\cdot, \cdot)$ is cosine similarity. The weighting function $w(t)$ is computed as:

$$w(t) = \text{sigmoid}(\log(\frac{\alpha_t}{\sigma_t + \epsilon})), \quad (7)$$

where α_t and σ_t are interpolant coefficients from the noise schedule, and ϵ constant prevents division by zero. This log-sigmoid formulation smoothly scales alignment strength, ensuring weaker constraints in early timesteps when noise dominates and stronger constraints as the latent representations converge toward structured audio features. The total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}}. \quad (8)$$

The alignment loss weight is set to $\lambda_{\text{align}} = 0.5$, balancing the flow matching loss and alignment objective.

Another challenge in representation alignment from the sequence-length mismatch between the latent representations and the pretrained audio encoder representations. Unlike the image domain, where spatial dimensions can be adjusted through interpolation without disrupting structural integrity, audio representations such as mel-spectrograms and waveform embeddings cannot be resized or interpolated without distorting essential spectral and temporal patterns. To address this, we explore three strategies:

- **Global Pooling:** Applies average pooling to both the latent representation and encoder output to produce one averaged representation. This simplifies alignment but may lose temporal granularity.
- **Interpolation-based Matching:** Interpolates the latent representations to match the encoder’s sequence length, preserving temporal structure and alignment.
- **Convolution-based Projection:** We introduce a projection mechanism using a 1D convolutional layer with a kernel size of 1 to reshape the latent representation into the same sequence length as the encoder embedding while preserving temporal dependencies:

$$h'_\phi(z_t) = \text{Conv1D}(h_\phi(z_t)), \quad (9)$$

where the 1D convolution adjusts N_z to match N_a , effectively bridges sequence-length differences, enabling precise feature alignment across timesteps.

By modulating alignment dynamically and resolving sequence-length mismatches, TRA ensures that latent representations evolve meaningfully throughout the synthesis process, enhancing both perceptual quality and temporal synchronization in video-to-audio generation.

5. Experiments

5.1. Experiment Setup

Datasets We train our model on the VGGSound [3] dataset, which comprises 174,337 videos, each lasting 10 seconds, following the official training split. We truncate all videos to 8 seconds and downsample them to 4 FPS, as done in prior work [22]. The VGGSound [3] dataset is a diverse dataset encompassing a broad range of audio-visual events, making it well-suited for training models that generalize across different acoustic environments. For evaluation, we use the original test split of VGGSound, containing 14,336 videos, which we also truncate to 8 seconds to match the training setup. To assess the generalization capability of our model beyond the training domain, we conduct a zero-shot evaluation on the Landscape dataset [16], which features 1,000 high-fidelity 10-second videos depicting natural environmental scenes. To maintain consistency, we truncate all videos to 8 seconds for evaluation.

Methods	FD↓	FAD↓	FID↓	IS↑	KL↓	Acc(%)↑	CLIP↑	MOS-Q↑	MOS-A↑	Infer(s)↓	Params(M)↓
SpecVQGAN (BMVC'21) [13]	1.25	5.65	22.77	15.76	7.25	57.20	6.37	2.52 ± 0.20	2.64 ± 0.24	3.58	308
Im2Wav (ICASSP'23) [27]	1.99	9.37	24.30	21.19	6.45	65.56	9.04	2.24 ± 0.23	2.81 ± 0.26	218.40	448
Diff-Foley (NeurIPS'23) [22]	3.34	4.71	10.55	56.67	6.05	93.92	8.92	2.87 ± 0.29	3.36 ± 0.29	3.00	860
See & Hear (CVPR'24) [34]	2.27	5.55	21.35	19.23	6.94	58.14	8.30	2.79 ± 0.23	2.84 ± 0.19	18.25	1099
FoleyCrafter (arXiv) [40]	0.99	2.45	12.07	42.06	5.67	83.54	13.47	3.18 ± 0.20	3.20 ± 0.20	2.96	1252
Frieren (NeurIPS'24) [31]	1.22	1.33	11.72	38.56	5.60	97.15	13.59	3.20 ± 0.27	3.43 ± 0.24	0.52	372
MDSGen (ICLR'25) [25]	1.04	1.34	11.83	30.37	6.03	96.52	11.41	3.05 ± 0.29	3.21 ± 0.24	1.56	142
TARO (Ours)	0.47	0.94	8.21	56.60	5.71	97.19	14.10	3.87 ± 0.24	4.16 ± 0.21	0.58	258

Table 1. **Quantitative comparisons of video-to-audio models on the VGGSound dataset.** We use colors to denote the **first**, **second** and **third** places respectively. †: MOS-AQ and MOS-AV scores are evaluated on samples from both VGGSound and Landscape datasets.

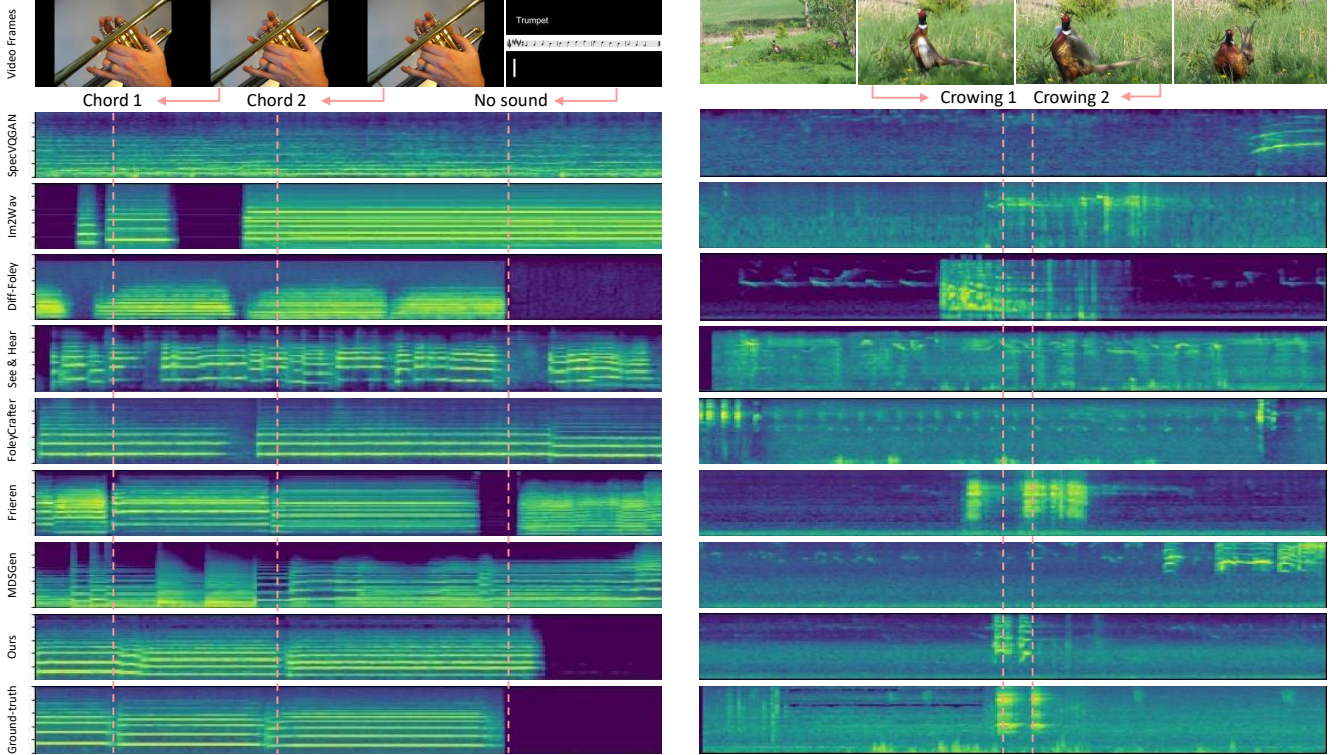


Figure 5. **Quantitative comparisons of video-to-audio models.** Our TARO achieves superior synchronization and fidelity, closely aligning with the ground truth, while other methods often produce misaligned or extraneous audio, underscoring its effectiveness in capturing event-driven acoustic details.

Model Configuration Our Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning (TARO) is a flow-based multimodal transformer with 12 transformer blocks and a hidden dimension of 768. We train the model for 500,000 steps on an NVIDIA A100 GPU using a batch size of 64 and a learning rate of $1e-4$, optimized with AdamW [21]. Only the Denoising Flow-based Transformer and Convolutional Projection are trained; other modules, such as the onset detection model, visual encoder, pretrained audio encoder, and VAE encoder/decoder, are frozen. To enable classifier-free guidance (CFG) [11], we randomly replace conditioning inputs with a zero vector 10% of the time during training, improving robustness

and generation quality. For audio processing, we leverage a frozen pretrained Variational Autoencoder (VAE) from AudioLDM2 [20] to encode and decode mel-spectrograms, paired with its vocoder for waveform generation. In our TRA strategy, we inject pretrained audio encoder knowledge into the 4th transformer blocks, employing EAT [6] as the audio encoder. We use a convolution-based projection to address sequence-length mismatches. For onset detection, we utilize a pretrained model from [8] to extract onset cues, processing 2-second segments from videos downsampled to 15FPS. During inference, we employ the SDE Euler-Maruyama sampler with 25 sampling steps and a CFG scale of 8.0 for optimal generation quality.

Methods	FD↓	FAD↓	Acc(%)↑	CLIP↑
SpecVQGAN (BMVC'21) [13]	2.71	8.23	74.80	6.55
Im2Wav (ICASSP'23) [27]	4.24	9.03	78.10	8.47
Diff-Foley (NeurIPS'23) [22]	7.99	5.55	84.40	8.75
See & Hear (CVPR'24) [34]	5.43	8.25	51.40	5.75
FoleyCrafter (arXiv) [40]	3.93	3.82	85.20	8.45
Frieren (NeurIPS'24) [31]	2.60	5.45	97.70	9.15
MDSGen (ICLR'25) [25]	3.80	6.13	97.40	10.53
TARO (Ours)	2.53	5.41	98.20	10.68

Table 2. **Quantitative comparisons of video-to-audio models on the Landscape dataset.**

Metrics We evaluate our audio generation model using a comprehensive set of metrics. For objective evaluations, we compute Frechet Distance (FD) and Frechet Audio Distance (FAD) from AudioLDM [19], alongside Frechet Inception Distance (FID), Inception Score (IS), and Mean KL Divergence (KL) using SpecVQGAN [13] scripts. Alignment Accuracy, based on scripts from prior works [22], quantifies audio-visual synchronization. Additionally, we measure perceptual similarity between video frames and generated audio using CLIP Score, computed with Wav2CLIP [32] for audio and CLIP [26] for video. We extract one video frame every 30 frames, average the visual features, and compute similarity with the corresponding audio features. The inference time is conducted on a single NVIDIA A100 GPU with a batch size of 1. For subjective evaluation, we perform a user study where participants rate generated audio samples using 1-5 Likert scales [17]. We report Mean Opinion Scores (MOS) for audio quality (MOS-Q) and content alignment (MOS-A), providing 95% confidence intervals. To ensure diversity in evaluation, we assess samples from VGGSound [3] and Landscape [16], covering a wide range of audio-visual scenarios. Additional details of evaluation are in *Suppl Sec. B*.

Baseline Models We evaluate our TARO against SOTA video-to-audio models, including SpecVQGAN [13], Im2Wav [27], Diff-Foley [22], See & Hear [34], FoleyCrafter [40], Frieren [31], and MDSGen [25]. For MDSGen, we select the Audio-VAE variant, as it offers significantly superior audio quality while maintaining high alignment accuracy compared to the Image-VAE version. We conduct both quantitative and qualitative evaluations to assess audio fidelity, semantic relevance, and temporal synchronization. The comparison provides a comprehensive analysis of how TARO performs relative to existing approaches across these key dimensions.

5.2. Main Results

Table 1 compares our TARO with SOTA video-to-audio methods [13, 22, 25, 27, 31, 34, 40] on the VGGSound dataset [3]. TARO outperforms all baselines in FD, FAD, and alignment accuracy, delivering high-fidelity, seman-

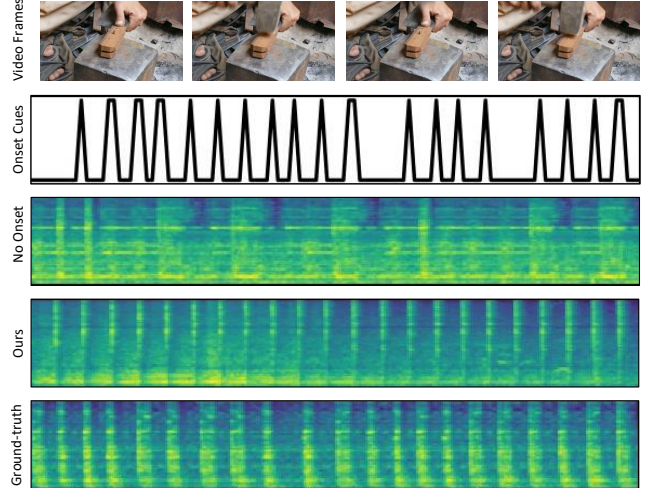


Figure 6. **Impact of Onset-Aware Conditioning on Audio Synchronization.** This figure compares generated spectrograms with and without OAC.

OAC	TRA	FD↓	FAD↓	Acc(%)↑	CLIP↑
		0.55	1.24	97.04	13.73
	✓	0.51	0.78	97.13	13.58
✓		0.53	1.10	97.16	14.05
✓	✓	0.47	0.94	97.19	14.10

Table 3. **Ablation study on key components of TARO.**

tically coherent, and well-synchronized audio. It also achieves the highest CLIP Score, indicating strong perceptual alignment with video content. Diffusion-based models like Diff-Foley [22] and FoleyCrafter [40] produce competitive results but suffer from high computational costs. Flow-based approaches such as Frieren [31] and masked diffusion models like MDSGen [25] improve efficiency but trade off quality and synchronization. In contrast, TARO balances performance across all metrics while maintaining fast inference (0.58s) and a compact model size (258M parameters), making it practical for real-time applications. Human evaluations further validate TARO’s effectiveness, achieving the highest Mean Opinion Scores with MOS-Q 3.87 and MOS-A 4.16, confirming its ability to generate perceptually natural, well-synchronized audio.

Fig. 5 presents the qualitative comparison of our TARO and SOTA video-to-audio methods [13, 22, 25, 27, 31, 34, 40]. TARO exhibits superior audio quality and synchronization, closely aligning generated spectrograms with ground-truth. On the right, it effectively reproduces two distinct crowing events in a pheasant video, ensuring seamless alignment with visual cues and enhancing perceptual coherence. These results highlight TARO’s ability to model fine-grained audio-visual correlations more effectively than prior methods. Additional audio samples and visualizations are provided in *Suppl Sec. F*.

	FD↓	FAD↓	Acc(%)↑	CLIP↑
w/o weighted	0.51	0.97	97.15	14.04
w/ weighted (Ours)	0.47	0.94	97.19	14.10
Global Pooling	0.51	1.16	97.06	14.01
Interpolate	0.49	1.09	97.13	14.12
Conv (Ours)	0.47	0.94	97.19	14.10

Table 4. **Ablation study on the details of Timestep-Adaptive Representation Alignment.**

Table 2 presents a quantitative comparison of TARO against state-of-the-art video-to-audio methods [13, 22, 25, 27, 31, 34, 40] on the Landscape dataset [16]. TARO achieves the highest synchronization accuracy and CLIP score, demonstrating superior alignment between generated audio and visual content. It also maintains competitive audio fidelity, closely matching the best-performing methods. These results highlight TARO’s strong generalization to diverse natural scenes, producing high-quality, well-synchronized audio without fine-tuning.

5.3. Ablation Studies

We conduct an ablation study on the VGGSound [3] test set to assess the impact of OAC and TRA in our TARO. Our full model integrates both components using a convolution-based projection with the EAT [6] audio encoder at the 4th transformer block.

As shown in Table 3, removing both components leads to the lowest scores. TRA alone improves distribution matching (FD, FAD), while OAC enhances synchronization (Acc, CLIP Score). Combining both achieves the best results, demonstrating their complementary effects on audio quality and alignment. Fig. 6 visualizes the impact of OAC. Without it, the generated spectrogram lacks a clear temporal structure and deviates from the ground truth. In contrast, incorporating OAC, our model closely aligns with the ground truth, ensuring precise synchronization between generated audio and visual events. Full metric evaluations for all ablation are provided in *Suppl Sec. C*.

Ablation on Timestep-Adaptive Representation Alignment To assess the impact of TRA, we compare weighting strategies and different sequence alignment methods in Table 4. Removing adaptive weighting reduces performance across all metrics, confirming its role in dynamically adjusting alignment strength. Global pooling weakens temporal granularity for sequence-length alignment, while interpolation improves CLIP Score but falls short in distribution matching. Our convolution-based projection achieves the best balance, preserving temporal dependencies and ensuring robust alignment. These results highlight the effectiveness of adaptive weighting and convolution-based alignment in enhancing audio fidelity and synchronization.

Ablation on Different Audio Feature Extractor We ex-

Encoder	FD↓	FAD↓	Acc(%)↑	CLIP↑
wav2vec 2.0 [2]	0.48	1.41	97.15	14.04
CLAP [33]	0.49	0.81	96.58	13.85
BEATs [5]	0.52	0.74	97.13	13.41
EAT [6] (Ours)	0.47	0.94	97.19	14.10

Table 5. **Ablation study on different audio encoders.**

Depth	FD↓	FAD↓	Acc(%)↑	CLIP↑
2	0.49	1.49	97.13	14.14
4 (Ours)	0.47	0.94	97.19	14.10
6	0.48	0.82	97.11	13.75
8	0.48	0.77	96.92	13.73
10	0.49	0.79	97.09	13.63

Table 6. **Ablation study on the effect of injecting audio encoder features at different transformer blocks.**

amine the effect of different pretrained audio encoders on TRA. Table 5 presents a comparison among wav2vec 2.0 [2], CLAP [33], BEATs [5], and EAT [6]. While all encoders contribute to strong performance, EAT consistently achieves the best results across FD, alignment accuracy, and CLIP Score while maintaining the lowest FAD. These results highlight EAT’s effectiveness in providing high-quality audio representations, making it the optimal choice for TARO by enhancing synchronization and fidelity in video-to-audio synthesis.

Ablation on Audio Encoder Injection Depth We analyze the effect of injecting audio encoder knowledge at different transformer blocks (2nd, 4th, 6th, 8th, and 10th) within TRA. As shown in Table 6, the 4th block achieves the best FD and alignment accuracy while maintaining a strong CLIP Score and competitive FAD. Injecting too early (2nd) or too late (6th, 8th, 10th) degrades performance, indicating that the 4th block offers the optimal balance between early integration and sufficient feature refinement, ensuring effective audio-visual synchronization without overfitting or underutilization of encoder features.

6. Conclusion

We present TARO, a novel Timestep-Adaptive Representation Alignment with Onset-Aware Conditioning framework for video-to-audio synthesis. By integrating Timestep-Adaptive Representation Alignment (TRA) and Onset-Aware Conditioning (OAC), TARO effectively aligns audio representations with pretrained audio encoders while capturing event-driven cues for precise synchronization. Our framework efficiently generates high-fidelity, temporally aligned audio from silent videos while maintaining computational efficiency. Benchmark evaluations demonstrate that TARO outperforms SOTA methods in fidelity, synchronization, and perceptual quality.

References

- [1] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014. 1
- [2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4, 8, 1
- [3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 2, 5, 7, 8, 1
- [4] Peihao Chen, Yang Zhang, Minghui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 2020. 1
- [5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv:2212.09058*, 2022. 3, 4, 8, 1
- [6] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. Eat: Self-supervised pre-training with efficient audio transformer. *arXiv:2401.03497*, 2024. 3, 4, 6, 8, 1, 2
- [7] Keunwoo Choi, Sangshin Oh, Minsung Kang, and Brian McFee. A proposal for foley sound synthesis challenge. *arXiv:2207.10760*, 2022. 1
- [8] Marco Comunità, Riccardo F Gramaccioni, Emiliano Postolache, Emanuele Rodolà, Danilo Comminiello, and Joshua D Reiss. Syncfusion: Multimodal onset-synchronized video-to-audio foley synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. 3, 4, 6
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 1, 3
- [10] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. 2
- [11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. 6
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3
- [13] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv:2110.08791*, 2021. 1, 2, 6, 7, 8
- [14] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2025. 2
- [15] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2
- [16] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic video generation. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 7, 8, 1
- [17] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932. 7, 1
- [18] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv:2210.02747*, 2022. 2, 3
- [19] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv:2301.12503*, 2023. 7
- [20] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 6
- [21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017. 6
- [22] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 3, 4, 5, 6, 7, 8
- [23] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024. 3
- [24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 1
- [25] Trung X Pham, Tri Ton, and Chang D Yoo. Mdsen: Fast and efficient masked diffusion temporal-aware transformers for open-domain sound generation. *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 6, 7, 8
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 7
- [27] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 1, 2, 6, 7, 8
- [28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 2
- [29] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue

- Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv:2312.15821*, 2023. [2](#)
- [30] Xihua Wang, Yuyue Wang, Yihan Wu, Ruihua Song, Xu Tan, Zehua Chen, Hongteng Xu, and Guodong Sui. Tiva: Time-aligned video-to-audio generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. [2](#)
- [31] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Conference on Neural Information Processing Systems (NeurIPS)*, 2025. [2](#), [6](#), [7](#), [8](#), [1](#)
- [32] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. [7](#)
- [33] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. [3](#), [4](#), [8](#), [1](#), [2](#)
- [34] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. [1](#), [2](#), [6](#), [7](#), [8](#)
- [35] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv:2408.06072*, 2024. [1](#)
- [36] Sunjae Yoon, Gwanhyeong Koo, Ji Woo Hong, and Chang D Yoo. Dni: Dilutional noise initialization for diffusion video editing. In *European Conference on Computer Vision (ECCV)*, 2024. [2](#)
- [37] Sunjae Yoon, Gwanhyeong Koo, Geonwoo Kim, and Chang D Yoo. Frag: Frequency adapting group for diffusion video editing. *arXiv:2406.06044*, 2024.
- [38] Sunjae Yoon, Gwanhyeong Koo, Younghwan Lee, and Chang Yoo. Tpc: Test-time procrustes calibration for diffusion-based human image animation. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024. [2](#)
- [39] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *International Conference on Learning Representations (ICLR)*, 2025. [1](#), [4](#)
- [40] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foley-crafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv:2407.01494*, 2024. [1](#), [2](#), [6](#), [7](#), [8](#)



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
✓	✓	0.47	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.

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 [13, 22, 25, 27, 31, 34, 40]. 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 [13, 22, 25, 27, 31, 34, 40] on both VGGSound [3] and Landscape [16] 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 [16]. 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

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 [33]	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.

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 [17], 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.

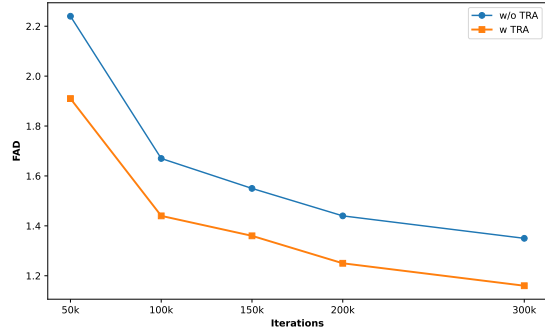


Figure 7. Training efficiency.

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 [33], 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 [14], down-sampled melspectrogram [30], and onset cues (Ours). We train all methods with 500k iterations. As shown, onset cues achieve the best balance, yielding

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.

the highest synchronization accuracy and CLIP score, with competitive FAD and FD.

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 [13, 22, 25, 27, 31, 34, 40]. 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.

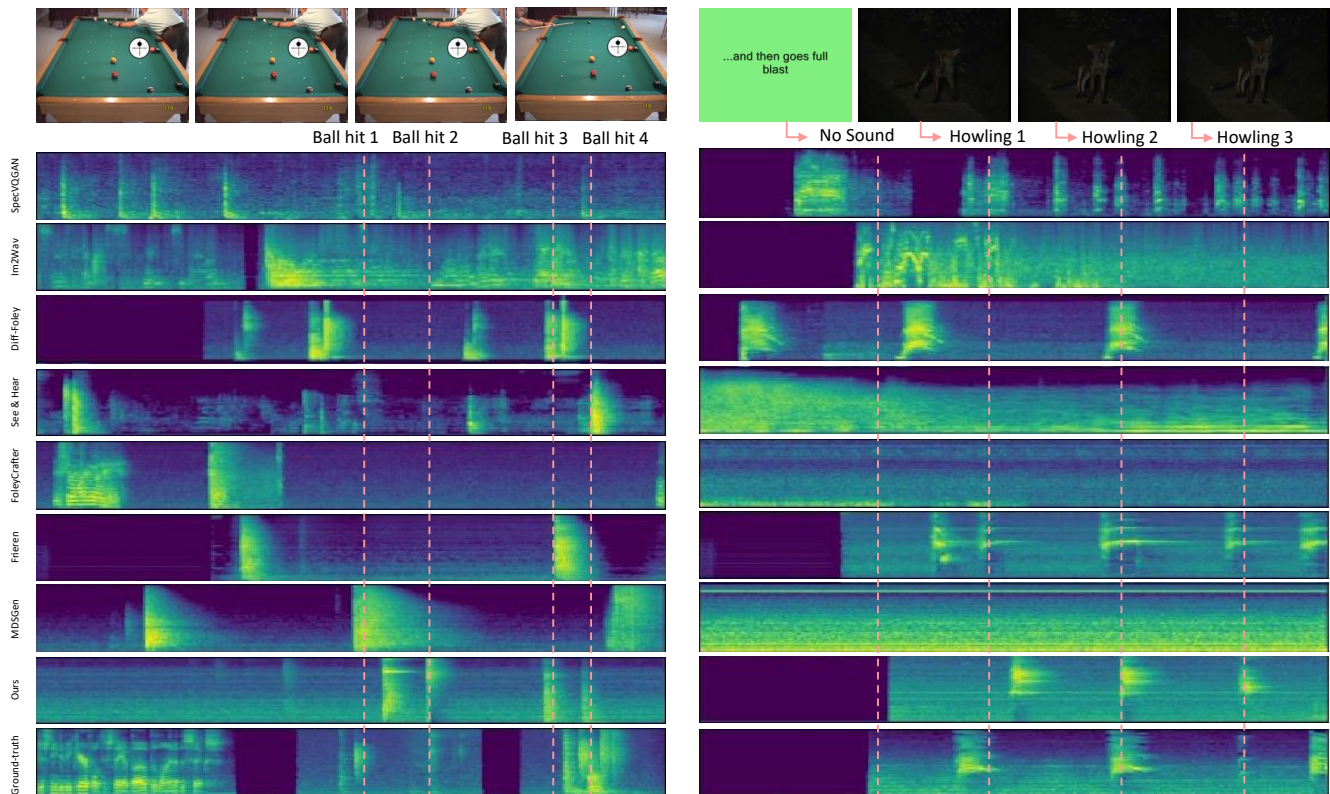


Figure 8. Visual comparisons on *1mMqLP36sCQ_000245* and *1JsIcP2nXMw_000108* from the VGGSound dataset.

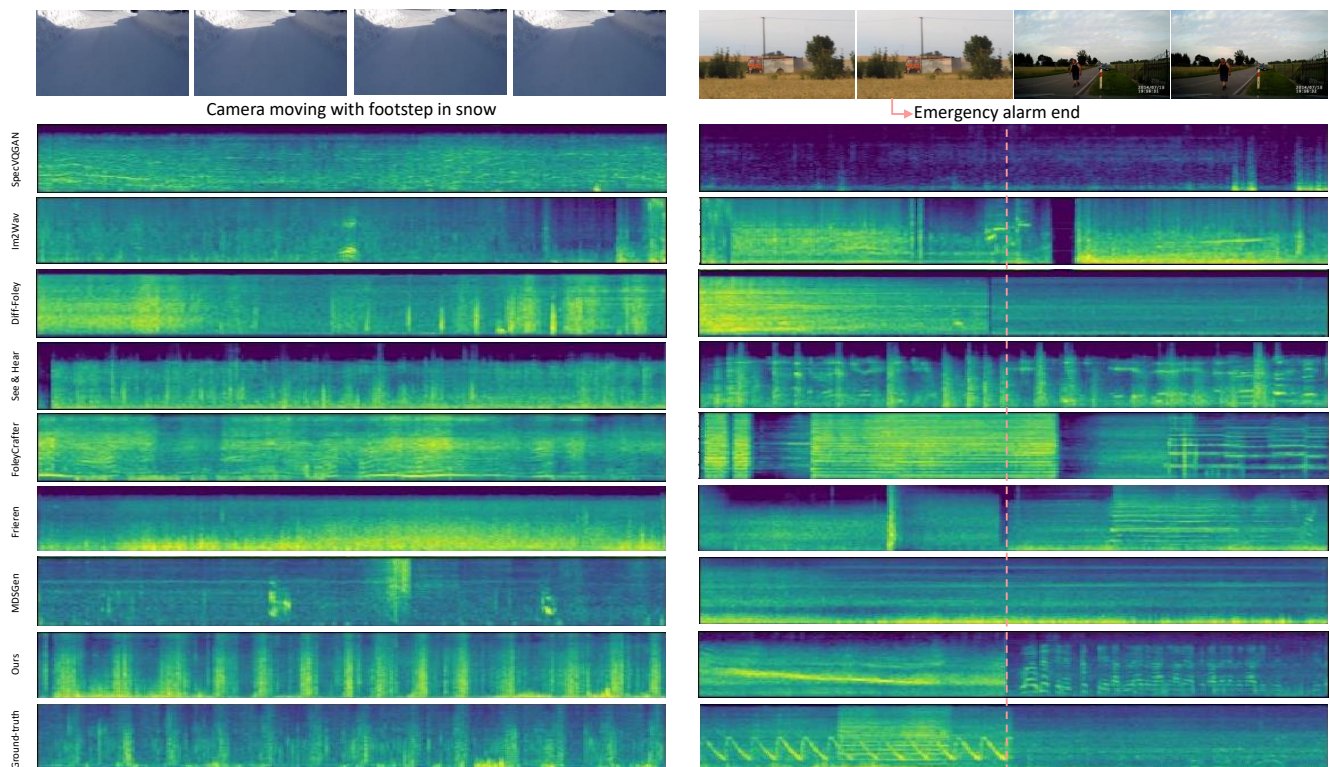


Figure 9. Visual comparisons on *KydSULgAHFI_000084* and *0N6S5OoG7Vg_000150* from the VGGSound dataset.

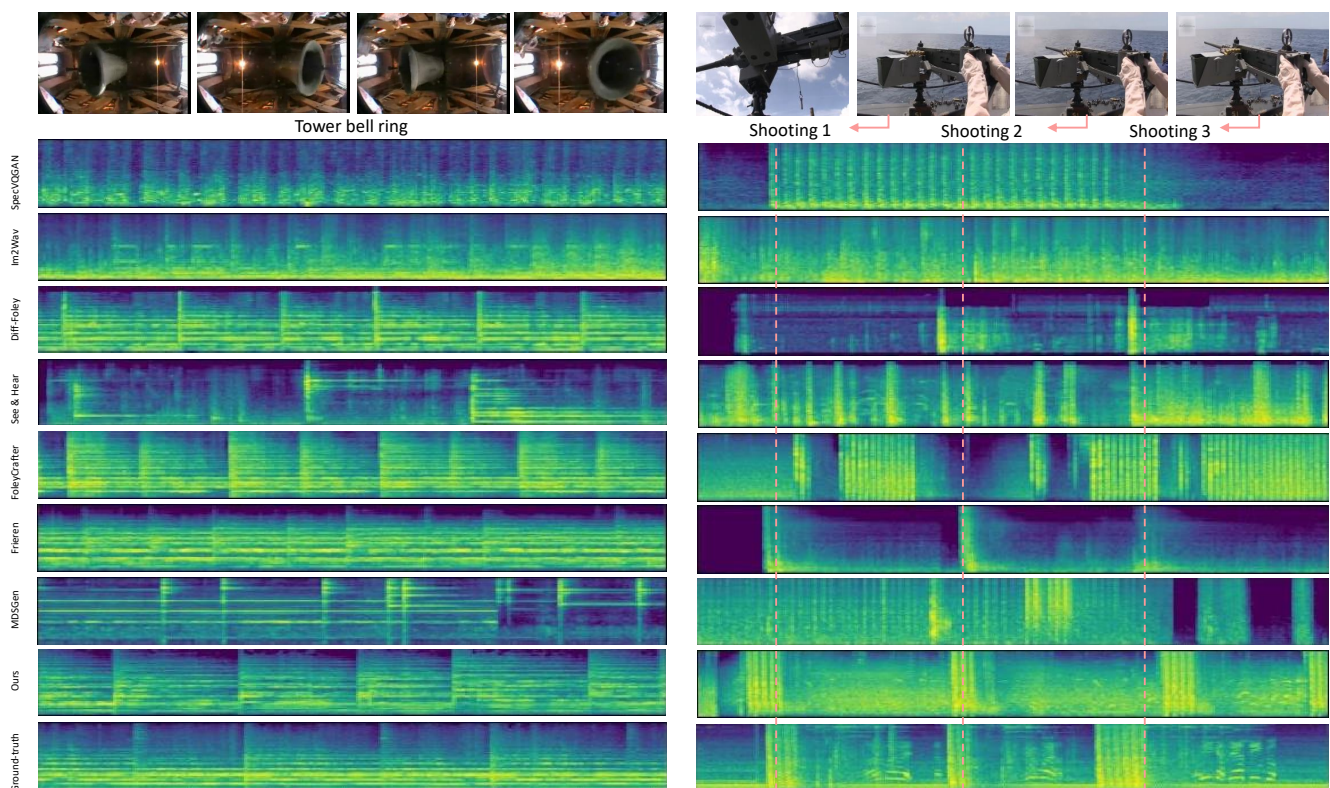


Figure 10. Visual comparisons on *6vl7eSBL-ag_000090* and *OFOHNgpDS38_000091* from the VGGSound dataset.