SOUND FOLLOWS CONTACT: STREAMING AUDIO-VISUAL GENERATION WITH CONTACT-AWARE STATE GROUNDING

Anonymous authors
Paper under double-blind review

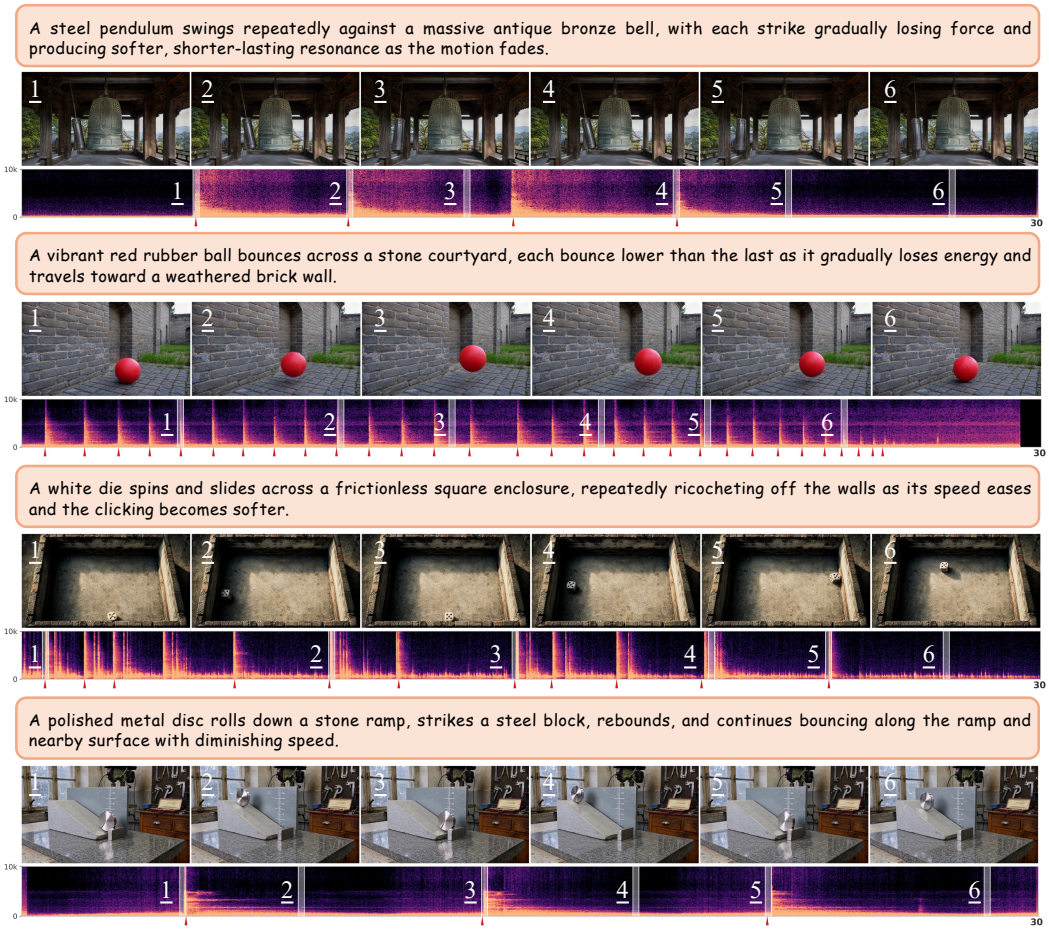

Figure 1: **Long-horizon 30 s streaming generations from Tora-live.** Rows show canonical contact scenarios: impact, bounce, swing, and rolling. Each row includes sampled frames and the full-length spectrogram. Generated audio stays silent during free flight, fires synchronously at visible contacts, and ceases once the object comes to rest. Red triangles mark contacts scheduled by state rollout. See the supplementary video for full rollouts.

ABSTRACT

Streaming audio-visual generation requires sound events to align with their corresponding visual causes. In rigid-body scenes, sound timing is strictly governed by physical contact, making it highly sensitive to subtle motion errors. Under autoregressive chunk-wise rollout, these errors accumulate across transitions and manifest as delayed, missing, or spurious audio events. We formalize this failure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mode as contact-phase drift, which compounds over long horizons where contact timing degrades substantially faster than perceptual quality. To address this challenge, we present **Tora-live**, a state-grounded framework for streaming audio-visual generation in contact-dominated rigid-body scenes. Tora-live augments a few-step chunk-causal generator with a compact contact-aware hybrid state that combines dominant-object kinematics with explicit contact-phase variables. This state is maintained online through lightweight state propagation, contact-phase construction, and closed-loop correction from generated video, conditioning both modalities simultaneously through current-state modulation and short-term state-history attention. We further introduce a contact-focused rollout optimization objective that encourages generated audio onsets to align with scheduled contacts while suppressing missed and spurious responses. To evaluate this setting, we build ContactBench, a benchmark for contact-dominated audio-visual generation with contact-centric alignment metrics. Experiments show that Tora-live improves contact timing, audio-visual alignment, and long-horizon robustness over strong streaming baselines while maintaining competitive perceptual quality, with gains that increase consistently with rollout length.

1 INTRODUCTION

Recent progress in joint audio-visual generation (Google DeepMind, 2025; OpenAI, 2025; Low et al., 2025; MiniMax, 2026; HaCohen et al., 2026) has enabled high-quality video synthesis with synchronized sound. However, streaming settings remain challenging (Su et al., 2026; Huang et al., 2026; Zhu et al., 2026), where audio and video must be generated autoregressively over long horizons. In this regime, self-conditioned rollout errors accumulate across chunk transitions, progressively degrading motion consistency and cross-modal alignment.

This challenge becomes particularly acute in contact-dominated rigid-body scenes (Zhang et al., 2026; Chen et al., 2025), where salient sound events are governed by discrete contacts such as impacts, rebounds, and sliding (Liao et al., 2026). For example, a falling ball should remain silent before impact, while a rolling object emits frictional sound only under sustained contact. Consequently, even subtle motion deviations can shift the underlying contact schedule, manifesting as delayed, missing, or spurious audio events during autoregressive rollout. We formalize this failure mode as contact-phase drift.

Existing streaming methods mainly improve stability through causal distillation (Yin et al., 2025; Li et al., 2026a), sliding-window attention (Huang et al., 2026; Su et al., 2026), or compressed cross-modal memory (Ding et al., 2026). While these mechanisms preserve local continuity and coarse synchronization, they do not explicitly model the contact phase that determines sound timing in rigid-body scenes. Consequently, generated videos often appear visually plausible while contact timing gradually drifts. Our experiments show that contact timing degrades substantially faster than visual perceptual quality as rollout length increases.

To address this problem, we introduce **Tora-live**, a state-grounded framework for audio-visual generation in streaming rigid-body scenes. Instead of relying on heavy physical simulation (Chen et al., 2025; Liu et al., 2026c), Tora-live maintains a compact contact-aware hybrid state for the dominant motion-bearing object, anchoring contact timing during streaming generation. Specifically, this state combines short-horizon kinematics with contact-phase variables. During rollout, it is propagated within each chunk by a lightweight dynamics prior, updated through contact-phase construction, corrected from generated video in a closed loop, and injected into the audio-visual backbone through current-state modulation and short-term state-history attention.

Beyond state grounding, we optimize the model via a contact-focused rollout objective that rewards generated audio onsets aligned with scheduled contacts while penalizing missed or spurious responses. To evaluate this setting, we establish ContactBench, a dedicated benchmark comprising 240 long-horizon real-world clips across six rigid-motion regimes: impact, bounce, contact transition, swing, rolling, and scraping. Each sample is comprehensively annotated with verified state trajectories and reference contact timestamps, supported by contact-centric temporal alignment metrics. Extensive experiments show that Tora-live improves contact timing, audio-visual alignment,

and long-horizon robustness across diverse rigid-body scenarios while maintaining competitive perceptual quality, with gains scaling consistently with rollout length.

Our contributions are summarized as follows:

- We identify contact-phase drift as a key failure mode in streaming rigid-body audio-visual generation and show that contact timing deteriorates faster than perceptual quality under long-horizon autoregressive rollout.
- We propose Tora-live, a state-grounded streaming framework that maintains and injects a compact contact-aware hybrid state to stabilize contact-timed audio-visual generation. Figure 1 shows representative 30-second generations, where contact-triggered sound events remain aligned with visible impacts, bounce, and rolling transitions over long horizons.
- We introduce a contact-focused rollout optimization objective that improves alignment between scheduled contact events and generated audio onsets under autoregressive streaming.
- We establish ContactBench, a benchmark comprising 240 long-horizon real-world clips across six contact regimes, equipped with verified annotations and contact-centric metrics to rigorously assess streaming audio-visual synchronization.

2 Related Work

Streaming audio-visual generation. Recent joint audio-visual generators, including Javis-DiT (Liu et al., 2025), Ovi (Low et al., 2025), LTX-2 (HaCohen et al., 2026), and MOVA (OpenMOSS Team et al., 2026), achieve strong perceptual quality but operate in an offline manner. Several streaming approaches extend causal distillation (Yin et al., 2025), self-forcing (Huang et al., 2026), and recurrent memory (Cui et al., 2025) to dual-stream generation (Su et al., 2026; Li et al., 2026a; Ding et al., 2026). These methods improve streaming stability by maintaining history context and cross-modal attention, but they do not explicitly model the contact phase that governs sound timing.

Structured and physics-aware generation. Another line of work guides video synthesis with structural conditions or explicit dynamics. In audio-visual generation, AVControl (Ben-Yosef et al., 2026), MMControl (Li et al., 2026b), MAVIN (Liu et al., 2026b), and Tora3 (Liao et al., 2026) introduce external guidance such as bounding boxes, depth maps, or motion trajectories. In visual generation, methods such as PhysGen3D (Chen et al., 2025), NewtonGen (Yuan et al., 2025), and RealWonder (Liu et al., 2026c) incorporate physics engines or numerical differential equation solvers to enforce dynamic plausibility. However, these simulation pipelines rely on heavy numerical integration, multi-stage particle mechanics, or iterative mesh tracking, which introduce substantial latency. Our setting instead requires a lightweight, contact-aware state representation that is inexpensive to propagate online and directly anchors sound emission timing.

3 Method

3.1 Overview

Tora-live builds on LTX-2.3 (HaCohen et al., 2026), which we distill into a four-step chunk-causal generator following (Su et al., 2026), denoted as LTX-2.3†. Generation proceeds in synchronized chunks: at chunk k, the model predicts video and audio latents (V_k, A_k) conditioned on previously generated chunks $\{(V_j, A_j)\}_{j<k}$. The central idea is to augment chunk-causal generation with a compact contact-aware hybrid state for the dominant motion-bearing object, instead of introducing full physical simulation into the streaming loop.

The framework consists of three parts: contact-aware state grounding, state-conditioned audio-visual generation, and contact-focused rollout optimization. Figure 2 summarizes the online inference pipeline: Tora-live maintains a compact contact-aware state throughout streaming inference and uses it to condition both modalities across chunk transitions. We further optimize the generator with a contact-focused rollout objective to improve realized contact timing under self-conditioned streaming.

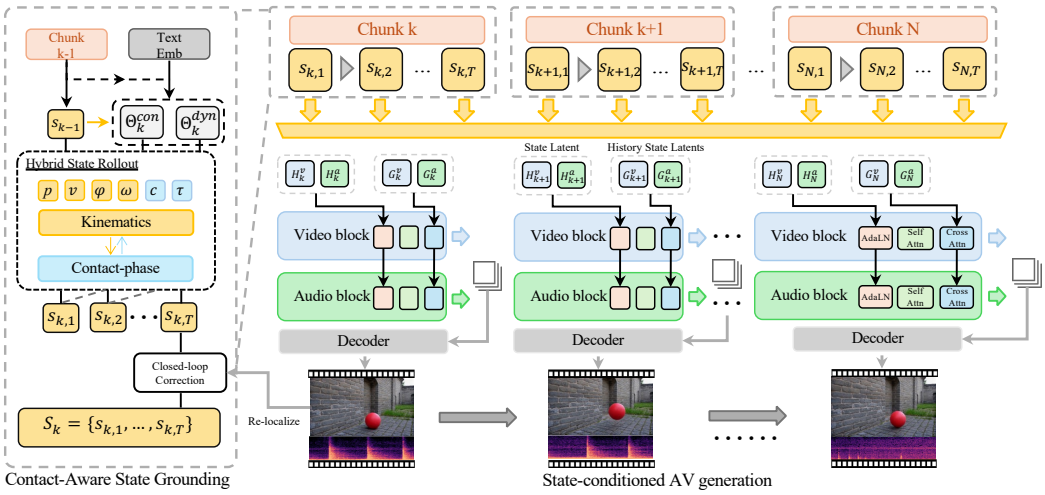

Figure 2: **Overview of Tora-live.** A compact contact-aware state is maintained online as a contact-timing anchor throughout streaming inference. Within each chunk, the state is rolled out under a lightweight dynamics prior together with contact-phase construction, then corrected from generated video before being passed to the next chunk. The aligned current state and short-term state history are used to condition both video and audio generation across chunk transitions.

3.2 Contact-Aware State Grounding

For chunk k, Tora-live maintains a state sequence

$$S_k = \{s_{k,1}, \dots, s_{k,T}\}, \qquad s_{k,t} = [p_{k,t}, v_{k,t}, \phi_{k,t}, \omega_{k,t}, c_{k,t}, \tau_{k,t}] \in \mathbb{R}^{10}, \tag{1}$$

where $p_{k,t}$ and $v_{k,t}$ denote position and translational velocity, $\phi_{k,t}$ and $\omega_{k,t}$ denote the in-plane principal orientation of the dominant motion-bearing object and its angular rate, and $(c_{k,t}, \tau_{k,t})$ encode contact phase through a backward-looking contact field and a forward time-to-contact. The state is maintained at frame rate, while only a small set of chunk-level parameters is predicted once per chunk:

$$\Theta_k^{\mathrm{dyn}} = (a_k, \lambda_k, \mu_k, \lambda_k^{\omega}), \qquad \Theta_k^{\mathrm{con}} = (n_k, \rho_k, e_k), \tag{2}$$

where Θ_k^{dyn} controls short-horizon state evolution and Θ_k^{con} defines a chunk-local contact proxy (Chen et al., 2025; Liu et al., 2026c). Here a_k is a chunk-level driving term, λ_k and λ_k^{ω} are translational and angular damping coefficients, and μ_k controls local reversion toward the chunk-initial state. The contact parameters (n_k, ρ_k) specify the support proxy, while e_k controls the restitution strength at scheduled contact.

Hybrid state rollout. Given the corrected state at the end of chunk $k-1$ and the predicted parameters $(\Theta_k^{\mathrm{dyn}}, \Theta_k^{\mathrm{con}})$, we roll out the state sequence of chunk k frame by frame. We first propagate the kinematic channels with a lightweight second-order dynamics prior (Rao & Yap, 1995):

$$\dot{p} = v, \qquad \dot{\phi} = \omega, \qquad \dot{v} = a_k - \lambda_k v - \mu_k\,(p - p_{k,0}), \qquad \dot{\omega} = -\mu_k\,(\phi - \phi_{k,0}) - \lambda_k^{\omega}\omega. \tag{3}$$

We then construct the contact phase from the propagated state. At each frame, we evaluate a signed clearance between the propagated body position and the support proxy. A sign change in this clearance between consecutive frames indicates a scheduled contact event. Based on the resulting event sequence, each frame is assigned a forward event clock $\tau_{k,t}$ that measures the time to the nearest scheduled contact ahead.

At the same time, we infer a frame-wise contact regime $r_{k,t} \in \{\textsc{free}, \textsc{sustain}, \textsc{rest}\}$ from the current clearance together with the normal and tangential components of the velocity. This regime determines whether the object is freely moving, in sustained contact, or at rest near contact. The backward contact field $c_{k,t}$ is then constructed from the most recent scheduled contact together with

the current regime, so that $\tau_{k,t}$ provides anticipatory timing and $c_{k,t}$ captures residual or ongoing contact effects.

Finally, when a scheduled contact is reached with an approaching normal velocity, we apply a contact-aware velocity update controlled by the restitution parameter e_k. As the normal approach speed decreases, the effective update strength is reduced, allowing repeated impacts to transition smoothly into sustained contact or rest.

Contact-phase construction is interleaved with frame-wise kinematic rollout, so the inferred contact regime can directly affect subsequent motion within the same chunk. Exact rollout details are given in Appendix A.2.

Closed-loop correction. After chunk k is generated, we decode its last frames, re-localize the dominant motion-bearing object, and extract an observed state $s_{k,T}^{\text{obs}}$ from the generated video. We then fuse its kinematic channels with those of the rolled-out state:

$$\bar{s}_{k,T} = (1 - \lambda_{\text{corr}})\, s_{k,T} + \lambda_{\text{corr}}\, s_{k,T}^{\text{obs}}, \tag{4}$$

where the fusion is applied only to the kinematic channels and $\lambda_{\text{corr}} \in [0, 1]$ controls the correction strength. This update re-anchors the state to realized visual evidence while preserving temporal continuity across chunks.

After correction, the contact-phase channels are rebuilt from the corrected kinematics. The corrected state at the end of chunk k initializes chunk $k + 1$, closing the streaming state loop.

3.3 State-Conditioned Audio-Visual Generation

Tora-live conditions both modalities on the contact-aware state. The current state sequence is aligned to the video and audio timelines separately and injected through two complementary mechanisms.

First, the current state modulates the latent features through modality-specific adaptive normalization:

$$(\gamma_m, \beta_m) = \Gamma_m(H_k^m), \qquad \hat{X}_k^m = \gamma_m \odot \text{Norm}(X_k^m) + \beta_m, \tag{5}$$

where $m \in \{v, a\}$, X_k^m denotes the latent features of the current chunk, and H_k^m denotes the aligned current state for modality m.

Second, the current chunk attends to the corrected states of the most recent L chunks:

$$\tilde{X}_k^m = \text{Attn}(\hat{X}_k^m, G_k^m), \tag{6}$$

where G_k^m denotes the aligned short-term state history for modality m. Current-state modulation provides chunk-local conditioning on the predicted contact phase, while state-history attention preserves recent motion and contact continuity across chunk transitions.

3.4 Contact-Focused Rollout Optimization

The state grounding module is first initialized on simulated rigid-body rollouts and then jointly trained with the audio-visual generator under the chunk-causal streaming protocol. Training details are deferred to Appendix A.3. Here we focus on the rollout-level objective that directly improves realized contact timing under self-conditioned streaming generation.

Figure 3 illustrates contact-focused rollout optimization. Let $\{t_i\}_{i=1}^{N_s}$ denote the scheduled contacts produced by the frozen state rollout and $\{\hat{t}_j\}_{j=1}^{N_o}$ the detected audio onsets in the generated output. After one-to-one matching within a tolerance window (Dixon & Spitz, 1980), we define the contact-focused timing reward

$$R_{\text{time}} = \frac{1}{N_s} \sum_{(i,j)\in P} \exp\Big(-\frac{|t_i - \hat{t}_j|}{\sigma_0}\Big) - \eta\Big(\frac{n_{\text{miss}}}{N_s} + \frac{n_{\text{extra}}}{N_o}\Big), \tag{7}$$

where $n_{\text{miss}} = N_s - |P|$ and $n_{\text{extra}} = N_o - |P|$ count missed and spurious onsets.

To preserve consistency between sound timing and the motion that should cause it, we add a rollout-level kinematic consistency term

$$R_{\text{kin}} = -\frac{1}{K}\sum_{k}\left\|s^{\text{obs}}_{k,T} - s_{k,T}\right\|_2^2, \quad (8)$$

where the norm is evaluated on the kinematic channels only. The total reward is

$$R = R_{\text{time}} + \alpha_{\text{kin}} R_{\text{kin}}. \quad (9)$$

We optimize this objective with Flow-GRPO (Liu et al., 2026a), where the frozen state grounding module provides the rollout-level contact schedule during self-conditioned training.

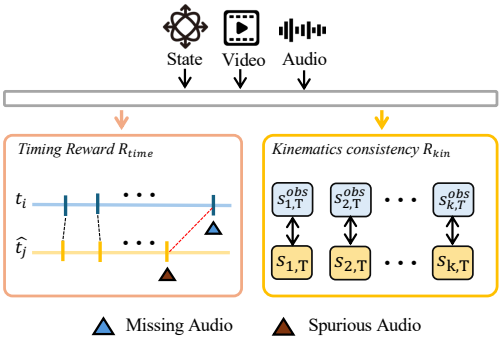

Figure 3: **Overview of contact-focused rollout optimization (CFRO).**

4 Experiment

4.1 Experimental Setup

Training data. We train Tora-live on a rigid-motion audio-visual dataset collected from VGGSound (Chen et al., 2020), ACAV-100M (Lee et al., 2021), and OpenVid1M (Nan et al., 2025). After filtering for rigid or object-centric motion, the dataset contains about 1.4M clips. We additionally use 60K clean rigid-body videos rendered with Genesis (Genesis Authors, 2024) for simulation-based state initialization.

Benchmarks. Our main evaluation is conducted on ContactBench, a benchmark of 240 clips covering six rigid-motion regimes. Detailed category statistics and benchmark construction are provided in Appendix B.2. We additionally report results on the physical-interaction subcategory of VABench (Hua et al., 2026) in Appendix D.3.

Metrics. We evaluate perceptual quality with Temporal Consistency (TC) (Zheng et al., 2025), Fréchet Video Distance (FVD) (Unterthiner et al., 2019), and Fréchet Audio Distance (FAD) (Kilgour et al., 2018). For audio-visual alignment, we report FGAS (Liao et al., 2026), defined as the cosine similarity between video frames and synchronized audio features in the latent space of CAV-MAE Sync (Araujo et al., 2025), together with three contact-centric metrics: Contact Timing Error (CTE), Missed Contact Rate (MCR), and Spurious Onset Rate (SOR). CTE measures the deviation between an audio onset and the visible contact it accompanies, while MCR and SOR quantify missed and spurious contact-triggered responses, respectively. FGAS captures fine-grained cross-modal feature alignment, whereas CTE, MCR, and SOR directly evaluate whether generated sound events are timed correctly with respect to visible contact. Runtime is measured on a single H100 GPU at 480P. Full metric definitions are provided in Appendix C.

4.2 Main Results

Table 1 shows that Tora-live improves contact-sensitive audio-visual alignment while preserving competitive perceptual quality. Compared with the streaming baseline LTX-2.3†, it consistently improves all contact-centric metrics, reducing CTE from 0.093 to 0.072, MCR from 0.232 to 0.182, and SOR from 0.189 to 0.158, while also improving FGAS from 0.216 to 0.243. In terms of generation quality and efficiency, Tora-live achieves the best TC and FAD, maintains competitive FVD, and incurs only a minimal latency overhead of 0.6 s over LTX-2.3†.

The cascaded V2A baseline illustrates that local timing correction alone is insufficient. Because MMAudio conditions on each decoded video chunk before synthesizing sound, it performs relatively well on local contact timing. However, independently synthesized audio chunks introduce splice artifacts and timbre drift, resulting in worse FAD and lower FGAS than both Tora-live and LTX-2.3†. Since the cascade leaves the video stream unchanged, it also cannot improve visual stability. By contrast, Tora-live improves both alignment and temporal consistency over the streaming baseline.

Table 1: **Comparison on ContactBench at 5 s.** LTX-2.3† denotes our streaming adaptation of LTX-2.3, trained on the same data and optimization settings as Tora-live, but without the proposed state grounding or CFRO. We use MMAudio-Large (Cheng et al., 2025) for dubbing. Best and second best per column are **bold** and underlined.

	AV alignment				Perceptual quality			Cost
Method	CTE↓	MCR↓	SOR↓	FGAS↑	TC↑	FVD↓	FAD↓	Runtime↓
Offline joint audio-visual models								
Ovi (Low et al., 2025)	0.144	0.406	0.334	0.185	0.798	301.7	2.52	134 s
MOVA (OpenMOSS Team et al., 2026)	0.117	0.295	0.257	0.214	0.826	258.3	2.15	837 s
LTX-2.3 (HaCohen et al., 2026)	0.108	0.269	0.254	0.211	<u>0.852</u>	**218.8**	1.94	90 s
Cascaded video to audio								
LTX-2.3† + MMAudio	<u>0.088</u>	<u>0.224</u>	0.200	0.212	0.836	234.2	2.24	10.2 s
Streaming models								
OmniForcing (Su et al., 2026)	0.158	0.418	0.428	0.179	0.782	328.6	2.92	**6.8 s**
LTX-2.3†	0.093	0.232	<u>0.189</u>	<u>0.216</u>	0.836	234.2	<u>1.83</u>	<u>7.4 s</u>
Tora-live	**0.072**	**0.182**	**0.158**	**0.243**	**0.861**	<u>221.5</u>	**1.71**	8.0 s

Table 2 shows that the advantage of Tora-live becomes larger as rollout length increases. From 5 s to 30 s, the contact-centric metrics of Tora-live deteriorate substantially more slowly than those of LTX-2.3†, and FGAS also decreases more gradually. The same trend appears when comparing with the offline teacher: although LTX-2.3† initially benefits from adaptation to rigid-motion data, this advantage is gradually eroded as rollout errors accumulate, and by 20 s the offline model overtakes it on both CTE and MCR.

Table 2: **Long-horizon behavior on ContactBench.** LTX-2.3 supports at most 20 s per generation.

Method	Metric	5 s	10 s	20 s	30 s
LTX-2.3 (offline)	CTE↓	0.108	0.114	0.127	–
	MCR↓	0.269	0.285	0.312	–
	SOR↓	0.254	0.267	0.291	–
	FGAS↑	0.211	0.207	0.201	–
	TC↑	0.852	0.847	0.832	–
LTX-2.3†	CTE↓	0.093	0.106	0.132	0.160
	MCR↓	0.232	0.271	0.338	0.392
	SOR↓	0.189	0.219	0.272	0.316
	FGAS↑	0.216	0.205	0.185	0.171
	TC↑	0.836	0.826	0.807	0.774
Tora-live	CTE↓	**0.072**	**0.081**	**0.099**	**0.121**
	MCR↓	**0.182**	**0.197**	**0.231**	**0.276**
	SOR↓	**0.158**	**0.167**	**0.198**	**0.234**
	FGAS↑	**0.243**	**0.235**	**0.222**	**0.212**
	TC↑	**0.861**	**0.853**	**0.836**	**0.816**

More importantly, contact-centric metrics deteriorate faster than perceptual quality in both streaming systems, showing that contact timing is the more fragile failure mode under long-horizon rollout. For LTX-2.3†, CTE, MCR, and SOR increase by 72%, 69%, and 67%, whereas temporal inconsistency $1 - \mathrm{TC}$ increases by only 38%. For Tora-live, the corresponding increases are reduced to 68%, 52%, and 48%, versus 32% for $1 - \mathrm{TC}$. This shows that Tora-live improves not only short-horizon alignment, but also the long-horizon stability of contact timing.

Figure 4 shows two representative 30-second comparisons. Even the offline teacher LTX-2.3 exhibits occasional long-range visual or timing errors, while the streaming baseline LTX-2.3† shows additional rollout-related failures such as premature or missing audio around visible contact. In contrast, Tora-live preserves more stable motion and more accurate contact-timed sound generation over long horizons.

4.3 Ablation Studies

All ablations are evaluated at 20 s. Row ⋆ in Table 3 is the reference configuration with the full contact-aware hybrid state, both injection mechanisms, and $\lambda_{\mathrm{corr}} = 0.3$, before contact-focused rollout optimization (CFRO). Relative to the baseline LTX-2.3†, state grounding alone already yields

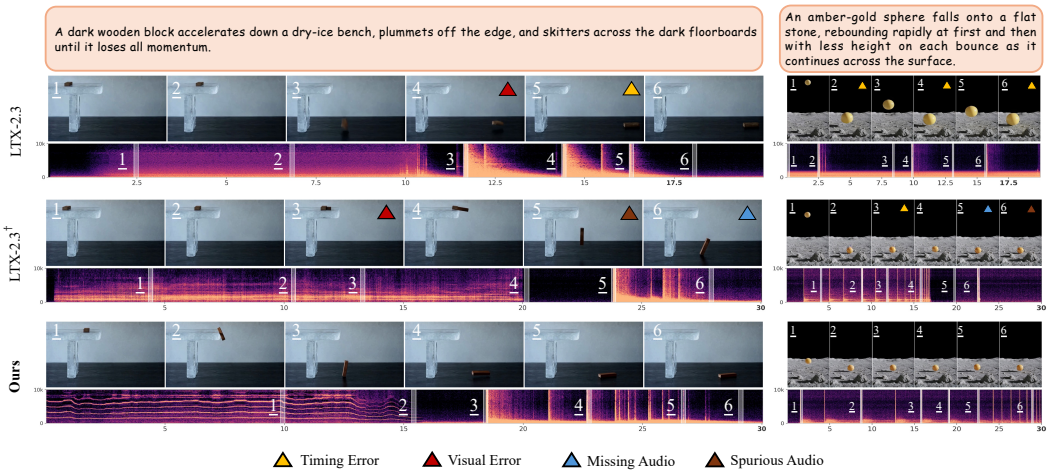

Figure 4: **Qualitative comparison on long-range contact-dominated scenes.** Examples show falling-sliding and bouncing dynamics with repeated contacts. Yellow, red, and blue markers denote timing error, visual error, missing audio and spurious audio at contact, respectively.

Table 3: **Ablations on ContactBench at 20 s.** Row $\star$ is the reference configuration (full hybrid state, both injection mechanisms, $\lambda_{\text{corr}} = 0.3$, no CFRO); each row in (a)–(d) changes one component of this reference.

	AV alignment				Perceptual quality		
Configuration	CTE↓	MCR↓	SOR↓	FGAS↑	TC↑	FVD↓	FAD↓
LTX-2.3† (baseline)	0.132	0.338	0.272	0.185	0.807	269.4	1.95
⋆Reference: state grounding only	0.112	0.286	0.214	0.211	0.826	254.7	1.85
(a) State content							
kinematic only (p, v, ϕ, ω)	0.129	0.324	0.268	0.189	0.810	265.5	1.93
kinematic + c	0.119	0.292	0.230	0.205	0.814	261.2	1.87
kinematic + τ	0.115	0.288	0.242	0.206	0.820	257.6	1.88
(b) State injection							
modulation only	0.114	0.284	0.278	0.195	0.798	273.9	1.96
history attention only	0.124	0.302	0.226	0.203	0.822	256.3	1.86
(c) Closed-loop correction							
open loop, $\lambda_{\text{corr}} = 0$	0.130	0.306	0.260	0.203	0.813	263.6	1.90
hard replacement, $\lambda_{\text{corr}} = 1$	0.124	0.290	0.234	0.197	0.805	271.4	1.93
(d) Training strategy							
w/o simulation initialization	0.117	0.296	0.222	0.208	0.822	256.9	1.86
+ CFRO, i.e. **Tora-live**	**0.099**	**0.231**	**0.198**	**0.222**	**0.836**	**245.8**	**1.79**

substantial gains in AV alignment while also improving TC, FVD, and FAD. This confirms that explicit contact-aware state grounding is the main source of improvement over the streaming baseline.

These gains do not come from kinematics alone. The kinematic-only variant remains close to the baseline, whereas adding either c or τ recovers much of the improvement. The full hybrid state performs best overall, indicating that recent-contact memory and future contact timing contribute complementary information.

The two state injection mechanisms also serve different purposes. Modulation mainly preserves chunk-local timing, while history attention better supports cross-chunk continuity. Using both gives the best balance between contact timing and perceptual quality.

A similar trade-off appears in the correction strategy. Open-loop propagation achieves better FGAS and perceptual quality than hard replacement, but performs worse on all three contact-centric met-

rics. This suggests that propagation preserves smoother continuity, whereas explicit correction is still needed to control timing drift. The fusion rule in Eq. 4 provides the best overall balance.

The training strategy matters as well. Removing simulation-based initialization degrades all metrics, while adding CFRO further improves contact-centric alignment, with the largest gain on MCR. This indicates that CFRO is particularly effective at recovering weak or missing contact-triggered responses under rollout.

Figure 5 visualizes the corresponding failure modes. The absence of history attention breaks cross-chunk continuity, producing delayed or spurious sound events. Without closed-loop correction, the predicted physical state drifts from realized motion to misalign impacts. Finally, removing CFRO leaves the generator vulnerable to premature or false collisions. The full model eliminates all three failure types, sustaining robust contact-timed synchronization through to rest.

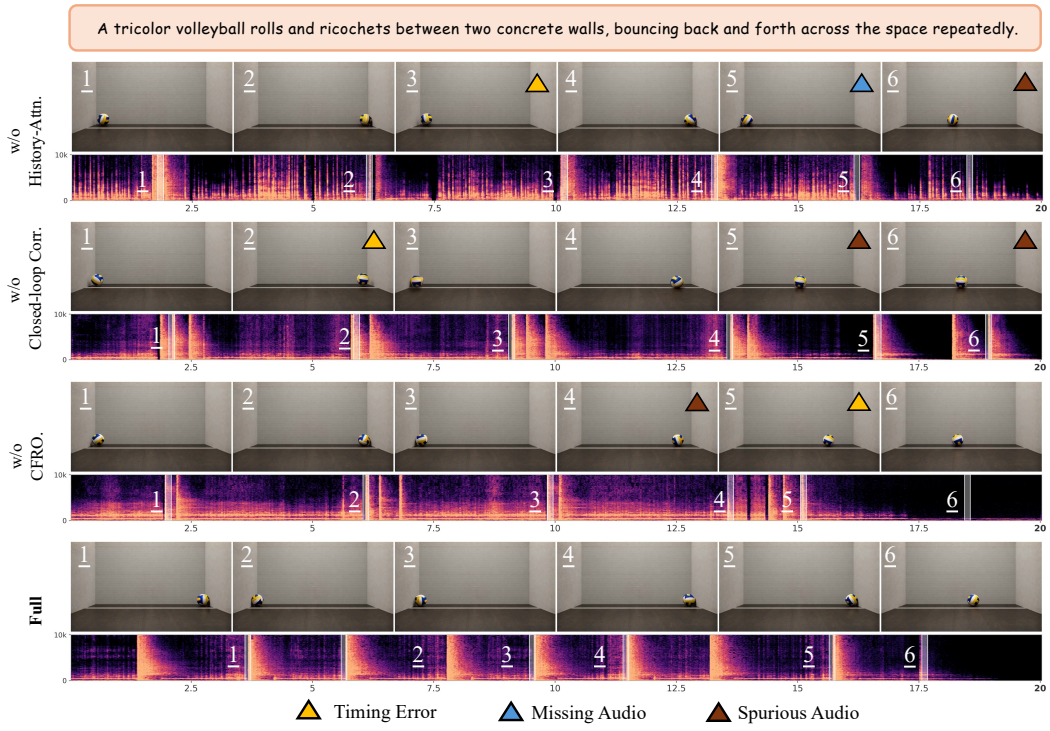

Figure 5: **Qualitative ablation on a long-range impact-and-rolling scene.** Removing history attention, closed-loop correction, or CFRO leads to distinct contact-timing failures. The full model preserves tighter alignment and cleaner decay to silence.

5 Conclusion

We presented Tora-live, a state-grounded framework for streaming audio-visual generation in contact-dominated rigid-body scenes. Our central finding is that long-horizon streaming failure is often driven by contact-phase drift, where self-conditioned rollout errors accumulate into delayed, missing, or spurious sound events around visible contacts. By maintaining a compact contact-aware hybrid state throughout streaming inference and optimizing the generator with contact-focused rollout supervision, Tora-live improves contact-centric audio-visual alignment and long-horizon robustness while preserving competitive perceptual quality. We also introduced ContactBench, a benchmark for contact-dominated streaming audio-visual generation with contact-centric alignment metrics. An important direction for future work is to move beyond the current dominant-object formulation toward multi-object relational state representations that can capture contact propagation across interacting bodies.

REFERENCES

Edson Araujo, Andrew Rouditchenko, Yuan Gong, Saurabhchand Bhati, Samuel Thomas, Brian Kingsbury, Leonid Karlinsky, Rogerio Feris, James R Glass, and Hilde Kuehne. Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18794–18803. IEEE, 2025.

Matan Ben-Yosef, Tavi Halperin, Naomi Ken Korem, Mohammad Salama, Harel Cain, Asaf Joseph, Anthony Chen, Urska Jelercic, and Ofir Bibi. Avcontrol: Efficient framework for training audio-visual controls, 2026. URL https://arxiv.org/abs/2603.24793.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6178–6189. IEEE, 2025.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 721–725. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053174. URL https://doi.org/10.1109/ICASSP40776.2020.9053174.

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*, pp. 28901–28911, 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation, 2025. URL https://arxiv.org/abs/2510.02283.

Yanbo Ding, Zhizhi Guo, Quanyue Song, Yishan He, Zhixiang He, Yongxiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory, 2026. URL https://arxiv.org/abs/2607.26818.

Norman F Dixon and Lydia Spitz. The detection of auditory visual desynchrony. *Perception*, 9(6): 719–721, 1980.

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

Google DeepMind. Veo 3. https://deepmind.google/models/veo/, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026. URL https://arxiv.org/abs/2601.03233.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

David N Lee. A theory of visual control of braking based on information about time-to-collision. *Perception*, 5(4):437–459, 1976.

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas M. Breuel, Gal Chechik, and Yale Song. ACAV100M: automatic curation of large-scale datasets for audio-visual video representation learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 10254–10264. IEEE, 2021. doi: 10.1109/ICCV48922.2021.01011. URL https://doi.org/10.1109/ICCV48922.2021.01011.

Chunyu Li, Jiaye Li, Ruiqiao Mei, Haoyuan Xia, Hao Zhu, Jingdong Wang, and Siyu Zhu. Hallo-live: Real-time streaming joint audio-video avatar generation with asynchronous dual-stream and human-centric preference distillation. *arXiv preprint arXiv:2604.23632*, 2026a.

Liyang Li, Wen Wang, Canyu Zhao, Tianjian Feng, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Mmcontrol: Unified multi-modal control for joint audio-video generation, 2026b. URL https://arxiv.org/abs/2604.19679.

Junchao Liao, Zhenghao Zhang, Xiangyu Meng, Litao Li, Ziying Zhang, Siyu Zhu, Long Qin, and Weizhi Wang. Tora3: Trajectory-guided audio-video generation with physical coherence. *arXiv preprint arXiv:2604.09057*, 2026.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

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.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. Mavin: Multi-shot audio-visual generation with customized narrative control, 2026b. URL https://arxiv.org/abs/2606.29473.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026c.

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

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, Oriol Nieto, et al. librosa: Audio and music signal analysis in python. *SciPy*, 2015(18-24):7, 2015.

MiniMax. Minimax h3. https://huggingface.co/MiniMaxAI/MiniMax-H3, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=j7kdXSrISM.

OpenAI. Sora 2. https://openai.com/index/sora-2/, 2025.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation, 2026. URL https://arxiv.org/abs/2602.08794.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Singiresu S Rao and Fook Fah Yap. *Mechanical vibrations*, volume 4. Addison-Wesley New York, 1995.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation, 2026. URL https://arxiv.org/abs/2603.11647.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025.

Qiyuan Zhang, Biao Gong, Shuai Tan, Zheng Zhang, Yujun Shen, Xing Zhu, Yuyuan Li, Kelu Yao, Chunhua Shen, and Changqing Zou. Physrvg: Physics-aware unified reinforcement learning for video generative models. *arXiv preprint arXiv:2601.11087*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Implementation Details

A.1 Streaming Baseline

We obtain LTX-2.3† by adapting LTX-2.3 (HaCohen et al., 2026) to four-step chunk-causal streaming generation following (Su et al., 2026). It operates on synchronized 1 s chunks, each containing 3 video latent frames and 25 audio latent frames. We first regress the causal student onto 3K ODE trajectories sampled from LTX-2.3 under the four-step schedule $\{1000, 757, 522, 0\}$, using AdamW with learning rate 2×10^{-4}, batch size 32, and 8K steps. We then distill the adapted model against the teacher with a distribution-matching objective, using learning rate 2×10^{-5} for the generator, 1×10^{-7} for the critic, five critic updates per generator update, batch size 32, gradient clipping 1.0, and 16K steps.

A.2 State Update under Streaming Rollout

The chunk-level parameter predictor is a lightweight two-layer transformer encoder. It takes four inputs: the clean video latent of the previous chunk, the object mask from the closed-loop correction step, the state sequence of the previous chunk, and the text embedding of the prompt. Each input is tokenized separately and projected to width 512. The video latent is flattened at full spatio-temporal resolution, and the object mask from the decoded tail frames is area-resampled onto the same spatial grid.

Initial state. At the start of inference, SAM3 (Carion et al., 2026) is prompted on the first frame to segment the dominant moving object. From the mask, we extract its centroid (u, v), projected area A, and principal orientation. Following the monocular size-based proxy in Lee (1976), we define the initial position channel as

$$p = \frac{1}{\sqrt{A}}(u, v, f_0), \tag{10}$$

where (u, v) is the centroid relative to the principal point and f_0 is a fixed depth-axis scale in pixels. This yields a scale-normalized 3D proxy in camera coordinates. Since p is normalized by $\sqrt{A}$, all state channels are dimensionless: positions are measured in units of the object's apparent diameter and velocities are computed by finite differences with $\Delta t = 1/24\,\mathrm{s}$. In practice, $\sqrt{A}$ is smoothed with a short moving average.

Support proxy and contact scheduling. Let $\bar{p}_{k-1,T}$ denote the position channel of the corrected state at the end of chunk $k-1$. For chunk k, the support offset is

$$d_k = \langle \bar{p}_{k-1,T}, n_k \rangle - \rho_k, \tag{11}$$

where (n_k, ρ_k, e_k) are the contact parameters in Θ_k^{con}. This defines a chunk-local contact boundary, and the signed clearance at frame t is

$$\chi_{k,t} = \langle p_{k,t}, n_k \rangle - d_k. \tag{12}$$

A sign change in $\chi_{k,t}$ between consecutive frames indicates a scheduled contact. The forward time-to-contact $\tau_{k,t}$ is then given by the nearest future scheduled event and clamped to $T_{\max} = 1.0\,\mathrm{s}$.

Contact regime and backward contact field. Let $v_{k,t}^n = \langle v_{k,t}, n_k \rangle$ and $v_{k,t}^{\parallel} = v_{k,t} - v_{k,t}^n n_k$ denote the normal and tangential velocity components. We normalize their magnitudes by a reference speed v_{ref}:

$$\tilde{v}_{k,t}^n = \frac{|v_{k,t}^n|}{v_{\mathrm{ref}}}, \qquad \tilde{v}_{k,t}^{\parallel} = \min\left(1,\ \frac{\|v_{k,t}^{\parallel}\|}{v_{\mathrm{ref}}}\right), \tag{13}$$

so that all velocity thresholds are dimensionless fractions of v_{ref}. For each frame, we assign a contact regime

$$r_{k,t} = \begin{cases} \textsc{sustain}, & |\chi_{k,t}| < \epsilon_h,\ \tilde{v}_{k,t}^n < \epsilon_n,\ \tilde{v}_{k,t}^{\parallel} > \epsilon_t, \\ \textsc{rest}, & |\chi_{k,t}| < \epsilon_h,\ \tilde{v}_{k,t}^n < \epsilon_n,\ \tilde{v}_{k,t}^{\parallel} \le \epsilon_t, \\ \textsc{free}, & \text{otherwise}, \end{cases} \tag{14}$$

where ϵ_h, ϵ_n, and ϵ_t are fixed thresholds. The backward contact field is

$$c_{k,t} = \max\Big(\exp(-t^-_{k,t}/T_0),\ \tilde{v}^{\parallel}_{k,t}\,\mathbf{1}[r_{k,t} = \textsc{sustain}]\Big)\,, \tag{15}$$

where $t^-_{k,t}$ is the elapsed time since the nearest past scheduled contact and T_0 is the decay constant. Here $\tau_{k,t}$ captures anticipatory timing, while $c_{k,t}$ captures residual or ongoing contact effects. In all experiments, we use $\epsilon_h = 0.10$, $\epsilon_n = 0.15$, $\epsilon_t = 0.40$, $T_0 = 0.15\,\mathrm{s}$, and $v_{\mathrm{ref}} = 2$. These values are calibrated on the simulated dataset.

Contact-aware velocity update. When a scheduled contact is reached with an approaching normal velocity , the translational velocity is updated by

$$v_{k,t} \leftarrow v_{k,t} - (1 + \hat{e}_{k,t})\, v^n_{k,t}\, n_k, \qquad \hat{e}_{k,t} = e_k \cdot \min\big(1,\ \tilde{v}^n_{k,t}\big)\,, \tag{16}$$

where e_k is the restitution parameter in Θ^{con}_k. As the normal approach speed decreases below v_{ref}, the effective restitution is attenuated. This allows repeated impacts to transition smoothly into sustained contact or rest instead of repeatedly triggering overly strong rebounds.

A.3 Training Configuration

Stage I. We initialize only the state grounding module on 60K simulated rigid-body clips rendered with Genesis (Genesis Authors, 2024). The simulator randomizes object shape, pose, velocity, support geometry, lighting, and texture, and provides exact short-horizon kinematics and contact events. Supervision is applied to both the predicted chunk-level parameters and the rolled-out state trajectories. No audio-visual backbone parameters are updated in this stage. The state grounding module is trained for 30K steps with learning rate 2×10^{-5} and batch size 64.

Stage II. We then jointly train the state grounding module and the audio-visual generator on real rigid-motion videos under the same training setting as the baseline. Training is conducted on 32 NVIDIA H100 GPUs. The rolled-out state is supervised with weak state targets extracted from the training videos, and the state grounding module is optimized with learning rate 2×10^{-6}. State propagation and correction follow the same procedure as at inference time. We add Gaussian perturbation with standard deviation 0.02 to the normalized state input and drop state channels with probability 0.1. The short-term state history is formed from the state latents of the previous $L = 2$ chunks, provided to the cross-attention module every four layers.

Stage III. For contact-focused rollout optimization, the state grounding module is frozen and used as a fixed contact-aware scheduler during self-conditioned generation. Unless otherwise specified, we use $\sigma_0 = 0.08\,\mathrm{s}$, $\eta = 0.3$, $\alpha_{\mathrm{kin}} = 0.2$, KL coefficient $\lambda_{\mathrm{KL}} = 0.1$, clipping parameter $\epsilon = 0.2$, and learning rate 1×10^{-6}. We draw 16 rollouts per prompt with the same first frame, clip the total reward to $[-2, 2]$, normalize advantages within each prompt group, and train for 300 steps.

B Dataset and Benchmark Details

B.1 Weak Target Extraction from Real Videos

For real videos, we extract weak state targets and contact events to supervise state-grounded training at scale. We first use Qwen3.5-Omni (Qwen Team, 2026) to identify the dominant motion-bearing object from a video-audio description, then use SAM3 to track it and extract its centroid, projected area, and principal orientation. Contact events are mined independently from video and audio, and only their intersection is retained.

On the visual side, we detect peaks in centroid acceleration magnitude using a prominence threshold of three median absolute deviations and a minimum separation of three frames. Direction reversals and clearance minima are added as auxiliary candidates. On the audio side, we detect onsets with librosa (McFee et al., 2015) at 24 kHz with hop length 512 and backtracking enabled. A visual candidate is retained only if an audio onset falls within ± 80 ms, and surviving events closer than 60 ms are merged.

We manually annotate 500 randomly sampled training clips and compare them with the automatically extracted tracks and contact events, counting an extracted event as correct when it falls within ± 0.2 s of an annotated contact. The pipeline identifies the dominant motion-bearing object in 88.4% of the clips and reaches 92.4% contact-event precision at 79.8% recall.

B.2 CONTACTBENCH

ContactBench contains 240 clips, each longer than 30 s, spanning six rigid-motion regimes: impact, bounce, contact transition, swing, rolling, and scraping. Each sample provides a prompt, a per-frame mask for the dominant motion-bearing object, a verified state trajectory, a set of reference contact times, and a category label. Rolling, scraping, and swing are not free of discrete contact events: most clips in these categories still contain intermittent or sustained contact within the evaluated window, often together with later impacts or contact transitions, and only a small subset contains no contact event in the window.

Table 4 reports per-category results together with the average number of annotated contact events per clip (1488 events in total). The largest gains over LTX-2.3† appear on impact, bounce, and contact transition.

Table 4: **Per-category results on ContactBench.** All values are clip-level means.

			LTX-2.3†			Tora-live		
Category	#	#ct/clip	CTE↓	MCR↓	SOR↓	CTE↓	MCR↓	SOR↓
Impact	48	8.5	0.078	0.204	0.162	**0.056**	**0.150**	**0.126**
Bounce	40	13.4	0.083	0.210	0.164	**0.060**	**0.154**	**0.130**
Contact transition	40	4.8	0.095	0.236	0.184	**0.066**	**0.169**	**0.142**
Swing	40	4.2	0.100	0.246	0.198	**0.078**	**0.192**	**0.166**
Rolling	40	2.6	0.102	0.248	0.214	**0.087**	**0.216**	**0.194**
Scraping	32	2.5	0.105	0.259	0.224	**0.092**	**0.226**	**0.206**
All	240	6.2	0.093	0.232	0.189	**0.072**	**0.182**	**0.158**

C CONTACT-CENTRIC METRICS AND HUMAN VALIDATION

C.1 METRIC DEFINITION

For evaluation, we use a stronger event extractor than the weak-label pipeline used for training. We first detect visual contact candidates and audio onset candidates from the generated sample, then query Qwen3.5-Omni (Qwen Team, 2026) to verify whether each candidate corresponds to a visible contact event or a salient audio onset. Nearby duplicates within 60 ms are merged.

Let the visible contacts be $\{t_i^{\text{vis}}\}_{i=1}^{N_c}$ and the detected audio onsets be $\{\hat{t}_j\}_{j=1}^{N_o}$. We assign onsets to visible contacts by the one-to-one matching

$$H = \arg\min_{P \in \mathcal{M}} \sum_{(i,j)\in P} \left|t_i^{\text{vis}} - \hat{t}_j\right|, \qquad N_{\text{hit}} = |H|, \tag{17}$$

where $\mathcal{M}$ is the set of maximal one-to-one matchings whose pairs satisfy $\left|t_i^{\text{vis}} - \hat{t}_j\right| \le \Delta_{\text{hit}} = 0.2$ s (Dixon & Spitz, 1980). We solve Eq. 17 with the Hungarian algorithm after masking out cost entries above Δ_{hit}.

The three contact-centric metrics are

$$\text{CTE} = \frac{1}{N_{\text{hit}}} \sum_{(i,j)\in H} \left|t_i^{\text{vis}} - \hat{t}_j\right|, \qquad \text{MCR} = 1 - \frac{N_{\text{hit}}}{N_c}, \qquad \text{SOR} = 1 - \frac{N_{\text{hit}}}{N_o}. \tag{18}$$

CTE measures timing deviation for matched events, MCR measures missed visible contacts, and SOR measures onsets without visible contact. All metrics are computed per clip and then averaged.

If no contact event is present in the reference video within the evaluation window and $N_{\text{c}} = N_{\text{o}} = 0$, we set $\text{CTE} = 0$ and $\text{MCR} = \text{SOR} = 0$. If $N_c = 0$ but $N_o > 0$, every onset is spurious and we set

Table 5: **Agreement between automatic metrics and human CTA preference.**

	FGAS	CTE	MCR	SOR
Agreement with CTA (%)	70.4	**87.2**	83.6	81.2

SOR $= 1$, excluding CTE and MCR from the average. If $N_o = 0$ but $N_c > 0$, we set MCR $= 1$ and CTE $= 0.2$ s, excluding SOR. Otherwise, if both sets are non-empty and no hit survives, we set CTE $= 0.2$ s and MCR $=$ SOR $= 1$.

C.2 Human Study

We compare Tora-live against LTX-2.3 and LTX-2.3† on 20 s generations. For each comparison, two samples produced from the same first frame and prompt are shown with model identities hidden and order randomized. Annotators choose *win*, *tie*, or *lose* for Tora-live under four criteria: visual quality (VQ), audio quality (AQ), contact timing alignment (CTA), and overall preference (OP).

CTA asks which sample better aligns audible events with visible contact timing. Each ContactBench prompt is evaluated against each baseline, giving 480 pairs, each rated by eight annotators and aggregated by majority vote. Figure 6 shows that the clearest advantage appears on CTA, while VQ, AQ, and OP also favor Tora-live.

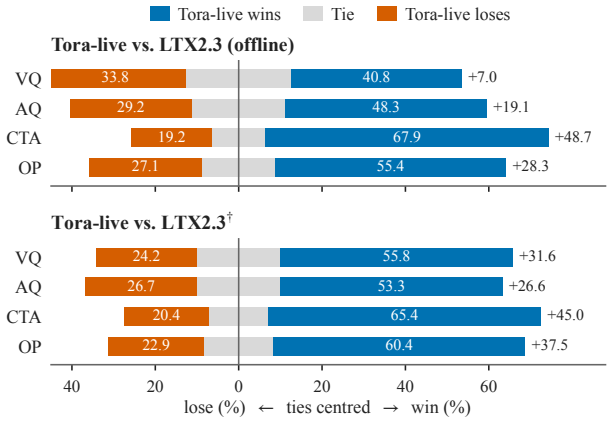

Figure 6: **Pairwise human evaluation at 20 s (%).**

We further compare automatic metrics with majority CTA judgments, excluding pairs whose majority vote is a tie. Table 5 shows that all three contact-centric metrics agree with human preference more often than FGAS, with CTE achieving the strongest agreement.

D Additional Analysis

D.1 Measurement Floor and Oracle State

The reported contact-centric metrics reflect both extraction noise and generation error. We separate these factors with two controls. First, we estimate a measurement floor by running the same metric pipeline on the ContactBench reference recordings, whose audio and video are synchronized by construction. Second, we evaluate an oracle-state setting by running Tora-live with the manually verified state trajectory, isolating the error attributable to imperfect state grounding.

Table 6 delineates the sources of timing error at 20 s. Providing oracle states improves CTE/MCR/SOR to $0.078/0.168/0.152$, establishing the performance ceiling achievable when state grounding is optimal. Comparing the oracle-state result against the measurement floor $(0.033/0.104/0.096)$ reveals that a considerable fraction of timing dispersion originates directly from the generative base model itself. Accurately modeling contacts therefore demands improvements across both state estimation and the foundational generator.

Table 6: **Error decomposition on ContactBench at 20 s.**

	CTE↓	MCR↓	SOR↓
Tora-live	0.099	0.231	0.198
Tora-live (with oracle state)	0.078	0.168	0.152
Reference annotations	0.033	0.104	0.096

Table 7: **What makes CFRO effective on ContactBench at 20 s.** SA denotes schedule alignment, defined as the fraction of scheduled contact events with an acoustic onset detected within 0.2 s. ⋆ Reference corresponds to the identical state-grounded configuration defined in Table 3.

Rollout objective	SA↑	CTE↓	MCR↓	SOR↓	TC↑	FVD↓	FAD↓
LTX-2.3†	-	0.132	0.338	0.272	0.807	269.4	1.95
LTX-2.3† + CFRO	-	0.124	0.335	0.240	0.805	270.0	1.93
⋆ Reference + CFRO w/o R_{kin}	**0.931**	0.118	0.292	0.242	0.818	262.4	1.90
⋆ Reference + CFRO	0.862	**0.099**	**0.231**	**0.198**	**0.836**	**245.8**	**1.79**

Table 8: **Results on physical-interaction subcategory of VABench.** Best and second best per column are **bold** and underlined.

	AV timing	Realism		Semantic quality	
Method	A-V Align↑	Visual↑	Audio↑	Visual QA↑	Audio QA↑
Offline joint audio-visual models					
Ovi	0.216	4.312	4.104	0.694	0.708
LTX-2.3	<u>0.259</u>	**4.571**	<u>4.252</u>	**0.758**	**0.784**
Streaming models					
OmniForcing	0.198	4.084	3.926	0.662	0.674
LTX-2.3†	0.231	4.458	4.176	0.736	0.761
Tora-live	**0.274**	<u>4.524</u>	**4.368**	<u>0.746</u>	<u>0.772</u>

D.2 What Makes CFRO Effective?

We test whether the gain of contact-focused rollout optimization (CFRO) comes from rollout optimization alone or from its combination with state grounding and kinematic consistency. We also monitor schedule alignment (SA), defined as the fraction of scheduled contact events with an acoustic onset detected within 0.2 s. Table 7 shows that CFRO alone yields limited gains. Applying CFRO directly to LTX-2.3† improves CTE slightly from 0.132 to 0.124 and reduces SOR from 0.272 to 0.240, while MCR remains nearly unchanged at 0.335 versus 0.338.

In contrast, CFRO with state grounding reaches $0.099/0.231/0.198$ on CTE/MCR/SOR. Crucially, removing R_{kin} inflates schedule alignment to 0.931 while severely degrading realized timing ($0.118/0.292/0.242$). This divergence exposes a reward-hacking effect, where the model fires onsets at scheduled timestamps irrespective of visible contact. Enforcing R_{kin} penalizes this decoupling, proving that rollout timing rewards require joint kinematic consistency to achieve true audio-visual synchronization.

D.3 Generalization to VABench

We additionally evaluate on VABench (Hua et al., 2026), using the physical-interaction subcategory with 47 cases in total. Table 8 reports the comparison using the official VABench metrics. Tora-live achieves the strongest A-V Align among the evaluated methods while remaining competitive on Visual Realism, Audio Realism, Audio QA, and Visual QA. This indicates that the benefit of contact-aware state grounding transfers beyond ContactBench.

D.4 More Qualitative Results

Figure 7 presents broader comparisons against all evaluated models on representative contact-dominated scenes. Figure 8 shows additional 30-second generations from Tora-live across diverse long-range rigid-body motions. We refer readers to the supplementary videos for full comparisons and complete rollouts.

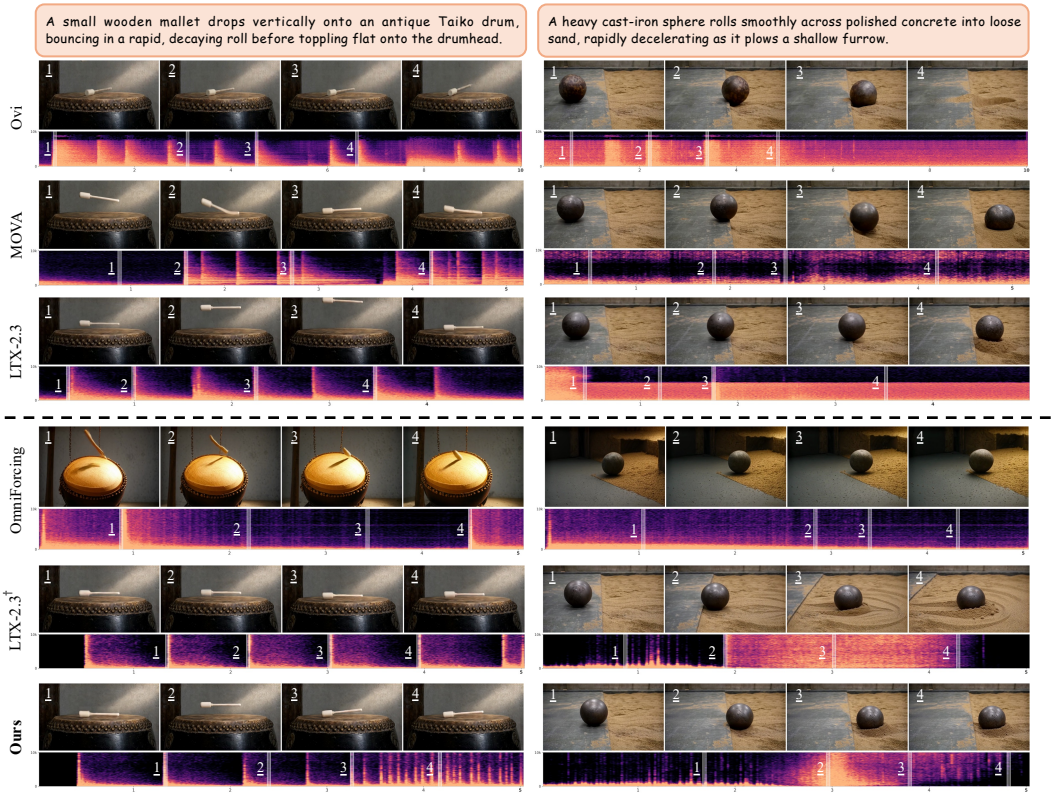

Figure 7: **Comparisons across all evaluated models on representative contact-dominated scenes.**

E Limitation

Tora-live uses a compact contact-aware state for the dominant motion-bearing object as a proxy for contact timing. This design is effective when the salient sound events are mainly governed by one dominant object, but it does not yet model full multi-object interactions.

As shown in Figure 9, Tora-live is less reliable in scenes such as a Newton's cradle, where motion and sound depend on sequential force transmission across interacting bodies. In such cases, a dominant-object state is insufficient to represent how contact timing propagates across the full object set. A natural next step is to extend the current formulation to a multi-object relational state, for example with interaction graphs or contact chains.

F AI USE STATEMENT

In preparing this manuscript, we used GPT-5.4 solely for language polishing and grammatical correction. The tool was used to improve clarity, readability, and fluency of the writing, but not to generate scientific ideas, formulate research claims, design methods, analyze results, write code, or make decisions about experiments or conclusions. All technical content, including the problem formulation, methodology, experiments, and interpretations, was developed and verified by the authors.

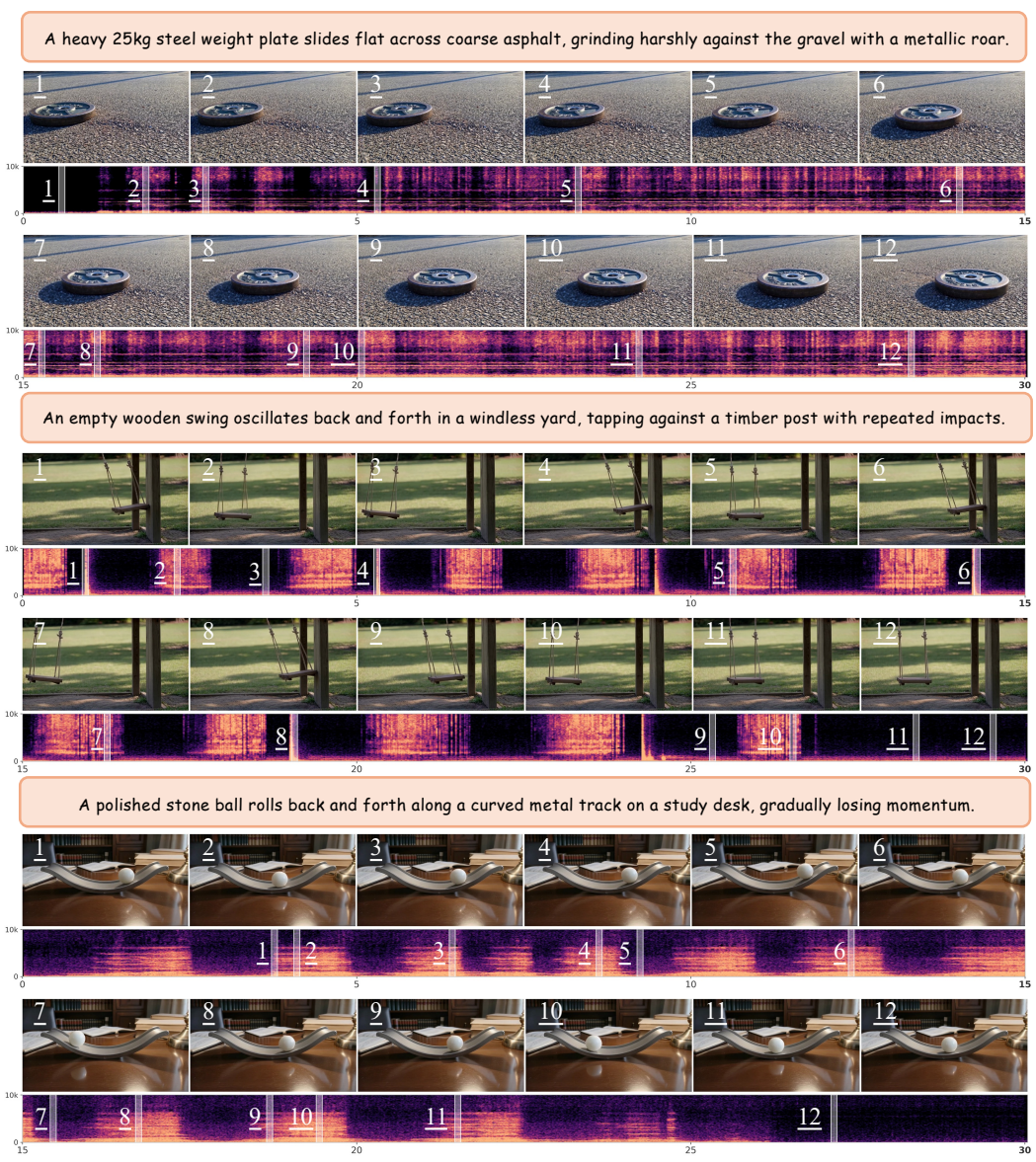

Figure 8: **Additional 30-Second generations from Tora-live on diverse long-range rigid-body motions.**

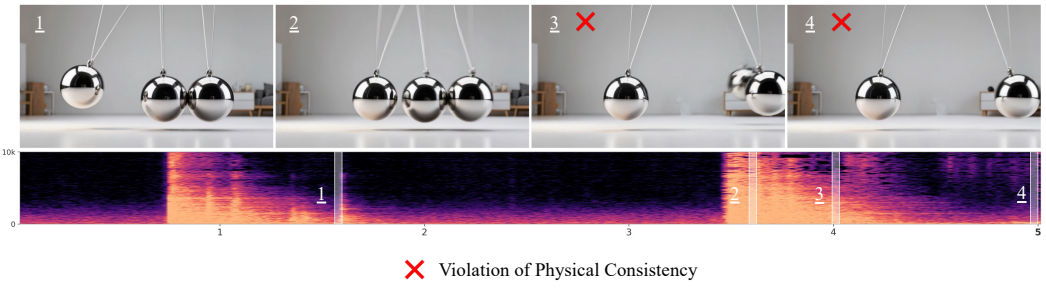

Figure 9: **Illustration of limitations in multi-object dynamic modeling.**

We reviewed and edited all AI-assisted revisions and take full responsibility for the final content of the paper.