OMNINFT: MODALITY-WISE OMNI DIFFUSION REINFORCEMENT FOR JOINT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

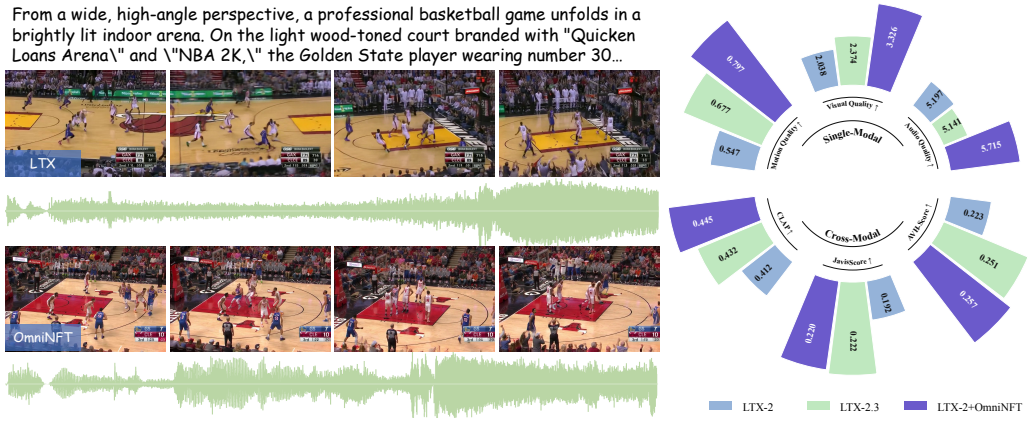

Figure 1: OmniNFT consistently improves the performance of LTX-2 in audio and visual quality, motion quality, cross-modal alignment, and audio–video synchronization.

ABSTRACT

Recent advances in joint audio-video generation have been remarkable, yet real-world applications demand strong per-modality fidelity, cross-modal alignment, and fine-grained synchronization. Reinforcement Learning (RL) offers a promising paradigm, but its extension to *multi-objective* and *multi-modal* joint audio-video generation remains unexplored. Notably, our in-depth analysis first reveals that the primary obstacles to applying RL in this stem from: **(i)** multi-objective advantages inconsistency, where the advantages of multimodal outputs are not always consistent within a group; **(ii)** multi-modal gradients imbalance, where video-branch gradients leak into shallow audio layers responsible for intra-modal generation; **(iii)** uniform credit assignment, where fine-grained cross-modal alignment regions fail to get efficient exploration. These shortcomings suggest that vanilla RL fine-tuning strategy with a single global advantage often leads to suboptimal results. To address these challenges, we propose **OmniNFT**, a novel modality-wise omni diffusion RL framework with three key innovations: **(1)** Modality-wise advantage routing, which routes independent per-reward advantages to their respective modality generation branches. **(2)** Layer-wise gradient surgery, which selectively detaches video-branch gradients on shallow audio layers while retaining those for cross-modal interaction layers. **(3)** Region-wise loss reweighting, which modulates policy optimization toward critical regions related to audio-video synchronization and fine-grained alignment. Extensive experiments on JavisBench and VBench with the LTX-2 backbone demonstrate that OmniNFT achieve comprehensive improvements in audio and video perceptual quality, cross-modal alignment, and audio-video synchronization.

1 INTRODUCTION

Recent years have witnessed significant advancements in joint audio-video generation (Low et al., 2025; Liu et al., 2026a; Seedance et al., 2025). However, achieving genuine practical utility in real-

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.

world applications demands a combination of high per-modality fidelity, robust cross-modal semantic consistency, and fine-grained audio-video synchronization. How to effectively align audio-video generative models with these multifaceted requirements remains an open and pressing challenge.

Despite rapid progress, current joint audio-video generative models (Wang et al., 2025; HaCohen et al., 2026; Liu et al., 2026a) still struggle to simultaneously satisfy these multifaceted objectives. Meanwhile, Reinforcement Learning with Verifiable Rewards (RLVR), particularly Group Relative Policy Optimization (GRPO) (Guo et al., 2025), has recently emerged as a powerful post-training paradigm for generative models (He et al., 2025; Li et al., 2025), owing to its ability to optimize complex and highly subjective objectives (Liu et al., 2026b). The previous research (Xue et al., 2025; Zheng et al., 2025) has successfully leveraged RLVR to enhance text-to-image/video generation quality and semantic alignment. These successes motivate a key question: ***can RLVR be effectively extended to joint audio-video generation to optimize fidelity, alignment, and AV synchronization?***

However, the performance of directly applying vanilla RLVR to joint audio-video generation remains suboptimal. Our in-depth analysis reveals that the primary obstacles stem from three types of optimization mismatch: **(i)** *multi-objective advantages inconsistency*: the different rewards for the video and audio components of a single multimodal output are often inconsistent, as illustrated in Fig. 2(a). **(ii)** *multi-modal gradients imbalance*: intra-modal generation and cross-modal interaction mainly concentrate in the shallow and deeper layers, respectively, see Fig. 2. However, gradients from the video branch tend to dominate the update direction of the shallow audio layers in the backward process. **(iii)** *uniform credit assignment*: AV synchronization and fine-grained alignment focus on certain critical regions, yet uniform updates ignore different contributions of these regions.

We address these challenges with **OmniNFT** (Modality-wise **Omni** Diffusion **N**egative-aware **F**ine-**T**uning), a novel modality-aware diffusion RL framework for joint audio-video generation. OmniNFT addresses the above issues with three coordinated designs. **(1) Modality-wise advantage routing:** for advantage inconsistency, we compute an independent advantage for each reward and selectively route it according to its underlying modality. **(2) Layer-wise gradient surgery:** for gradients imbalance, we detach the part gradient from the video stream on the shallow layers of the audio model, while preserving the effective gradients responsible for audio–video interaction. **(3) Region-wise loss reweighting:** for credit assignment, we introduce a critical-region reweighting strategy that strengthens the optimization on critical regions. This fine-grained credit assignment, coupled with gradient surgery, effectively circumvents optimization mismatch across modalities.

We conduct extensive experiments on JavisBench (Liu et al., 2025c) and VBench (Huang et al., 2024) using LTX-2 (HaCohen et al., 2026) as the backbone. Compared with strong baselines, OmniNFT achieves comprehensive improvements across audio and video perceptual quality, cross-modal alignment, and audio-video synchronization. Our contributions are summarized as follows:

- We first visit RL for joint audio-video generation and identify its core optimization bottleneck, including advantage inconsistency, gradients imbalance, and uniform credit assignment, which together hinder effective joint improvement.
- Based on these bottlenecks, we propose OmniNFT that combines modality-wise advantage routing, layer-wise gradient surgery, and region-wise loss reweighting, effectively facilitating multi-objective and multi-modal optimization.
- Extensive experiments demonstrate that OmniNFT delivers consistent gains in perceptual quality, cross-modal alignment, and audio-video synchronization over strong baselines.

2 Related Work

2.1 Text-to-Video Generation.

Text-to-video generation has rapidly progressed from U-Net-based diffusion to large-scale Transformer-based diffusion. Early methods (Blattmann et al., 2023a; Guo et al., 2023; Wu et al., 2023a; Blattmann et al., 2023b) typically adapt image diffusion backbones with temporal modules to support video generation, while recent Diffusion Transformers (DiTs) (Brooks et al., 2024; Hong et al., 2022; Kong et al., 2024) scale more effectively with data and model size, demonstrating substantial gains in visual fidelity, physical plausibility, and text alignment. Among open-source lines,

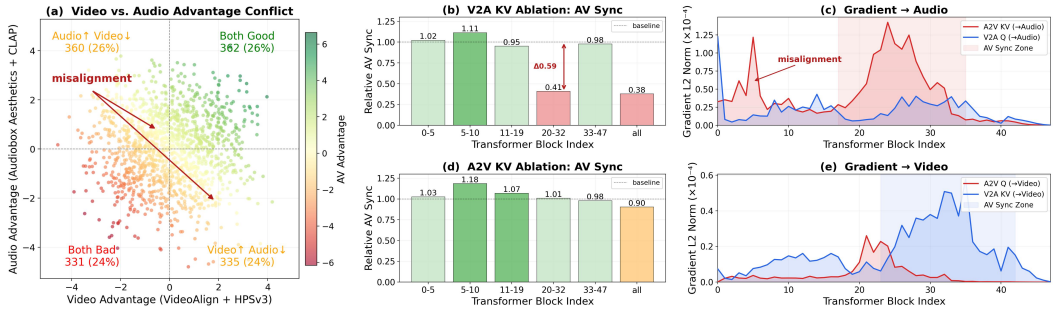

Figure 2: **Advantage inconsistency and asymmetric audio-video interaction.** (a) Video and audio advantages are weakly correlated: roughly half of the samples receive *opposing* rewards across the two modalities. (b) Blocking the V2A KV in mid-layers collapses AV synchronization to $0.41\times$ baseline, whereas (d) the symmetric A2V ablation causes a mild degradation when applied to the later blocks. (c,e) Layer-wise gradient norms of cross-attention show that audio-video interaction is concentrated in the *middle* and *later* transformer blocks (AV Sync Zone). On the audio branch, gradients from video KV (A2V) disturb the update direction of the audio shallow blocks in (c).

Wan (Wan et al., 2025) emphasizes efficient spatiotemporal attention with flow-matching training, and LTX-Video (HaCohen et al., 2024) further targets real-time inference by jointly optimizing a Video-VAE and a denoising Transformer.

2.2 Joint Audio-Video Generation.

Building on strong text-to-video backbones, recent work increasingly focuses on generating temporally synchronized video and audio in a single pipeline. Veo3 (Google DeepMind, 2024) demonstrates joint audio-video capabilities. UniVerse-1 (Wang et al., 2025) combines Wan 2.1 (Wan et al., 2025) for video generation and ACE-Step (Gong et al., 2025) for audio generation, and introduces lightweight projection modules to align latent spaces across modalities. In parallel, the Javis series (Liu et al., 2025c; 2026a) explores unified modeling for joint generation, whereas the LTX series (HaCohen et al., 2026) adopts an asymmetric dual-stream design with bidirectional cross-attention to handle modality imbalance. While these efforts have yielded promising results, there remains substantial room for improvement in generation quality and audio-visual synchronization.

2.3 Reinforcement Learning for Visual Generation.

Inspired by the success of GRPO (Guo et al., 2025) in large language models (LLMs), RLVR has emerged as a practical post-training strategy for improving visual quality and semantic alignment in visual generation (Luo et al., 2025; Zhang et al., 2025a;b; Jiang et al., 2025). Several methods adapt reward-driven policy optimization to image and video generation tasks. For example, T2I-R1 (Jiang et al., 2025) introduces dual-level chain-of-thought RL for autoregressive generation, while Flow-GRPO (Liu et al., 2025a) and Dance-GRPO (Xue et al., 2025) extend flow matching with stochastic trajectories by reformulating Ordinary Differential Equation (ODE) as Stochastic Differential Equation (SDE) processes. DiffusionNFT (Zheng et al., 2025) instead finetunes diffusion models in the forward process through an implicit policy-improvement direction. However, effective expansion of RL into multi-objective joint audio-visual generation remains under-explored.

3 Preliminary

Joint Audio-Video Flow Matching. In joint audio-video generation, an audio latent x^a and a video latent x^v are denoised in parallel under a shared flow matching (Lipman et al., 2022) schedule. Each modality is independently perturbed by its own Gaussian prior $x_1^m \sim \mathcal{N}(0, I)$, while sharing a timestep $t \in [0, 1]$:

$$x_t^m = (1-t)\, x_0^m + t\, x_1^m, \quad m \in \{a, v\}. \tag{1}$$

A dual-stream model $v_\theta = (v_\theta^a, v_\theta^v)$ jointly predicts the velocity of both modalities, where the two streams interact through cross-modal attention layers. During the sampling phase, both streams are integrated in parallel via a deterministic ODE solver:

$$dx_t^m = v_\theta^m(x_t^a, x_t^v, t, c)\, dt, \quad m \in \{a, v\}, \tag{2}$$

where c denotes the text condition.

Diffusion Negative-aware Finetuning (DiffusionNFT). In contrast to existing diffusion-based GRPO frameworks (Liu et al., 2025a; Xue et al., 2025), which typically necessitate a transition from deterministic ODEs to SDEs, DiffusionNFT (Zheng et al., 2025) performs policy optimization directly on the forward diffusion process. The method leverages a reward $r(x_0, c)$ to determine positive and negative policies, thereby defining a contrastive loss. The model's velocity predictor, v_θ, is encouraged toward the high-reward policy and away from the low-reward one. The core policy optimization loss is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{c,\pi^{old}(x_0|c),t}\Big[r\|v_\theta^+(x_t, c, t) - v\|_2^2 + (1-r)\|v_\theta^-(x_t, c, t) - v\|_2^2\Big], \tag{3}$$

The implicit positive and negative policies v_θ^+ and v_θ^- are combinations of the old policy v^{old} and the training policy v_θ, weighted by a hyperparameter β:

$$v_\theta^+(x_t, c, t) := (1-\beta)v^{old}(x_t, c, t) + \beta v_\theta(x_t, c, t), \tag{4}$$

$$v_\theta^-(x_t, c, t) := (1+\beta)v^{old}(x_t, c, t) - \beta v_\theta(x_t, c, t). \tag{5}$$

The optimality probability $r \in [0, 1]$ is transformed from the unconstrained raw reward signal r^{raw}:

$$r(x_0, c) := \frac{1}{2} + \frac{1}{2}\text{clip}\left[A(x_0, c), -1, 1\right], \quad A(x_0, c) = \frac{r^{raw}(x_0, c) - \mathbb{E}_{\pi^{old}(\cdot|c)} r^{raw}(x_0, c)}{Z_c}, \tag{6}$$

where $A(x_0, c)$ denotes the group-wise normalized advantage of sample x_0 under prompt c, $Z_c > 0$ is a normalizing factor, which could take the form of a global reward std.

4 Motivation

4.1 Advantage Inconsistency between Audio and Video

Joint audio-video generation requires high quality in both modalities, yet their reward feedback can differ for the same output. To examine this inconsistency, we analyze 1,400 generated samples from 175 prompts with a group size of 8 and compute video and audio advantages separately. As shown in Fig. 2(a), *nearly half of the samples exhibit opposite-signed video and audio advantages*, indicating that a rollout can outperform its group baseline in one modality while underperforming in the other. Fig. 3 illustrates the implications using two per-sample examples, with equal-weight aggregation and $A_{\text{av}} = 0$ for clarity. For $(A_v, A_a) = (-0.4, +0.4)$, shared advantage aggregation assigns $(0, 0)$ to the video and audio losses, canceling the opposing modality-specific preferences. For $(A_v, A_a) = (+0.1, -0.2)$, it assigns $(-0.1, -0.1)$, masking the positive video advantage. In contrast, modality-wise routing preserves the separate video and audio advantages of each sample rather than assigning the same aggregated advantage to both modality losses. In the OmniNFT, the cross-modal alignment advantage is added to both modality-specific advantages, preserving shared alignment supervision alongside modality-specific feedback.

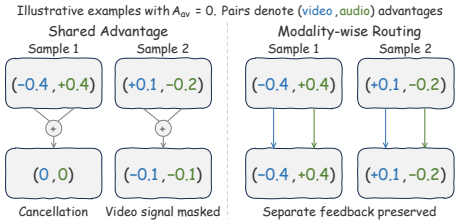

Figure 3: Shared advantage aggregation versus modality-wise routing.

4.2 Gradients Imbalance across Different Modalities Branches

The advantage inconsistency discussed above only reflects the issue in the output. We examine the internal information and gradients flow within the dual-stream model from both **forward** and **backward** perspectives, as illustrated in Fig. 2.

Forward Analysis. We identify the functional roles of the Transformer blocks at different depths by selectively blocking the key-value (KV) information exchanged through audio-video cross-attention. As shown in Fig. 2(b) and (d), ablating the KV flow in shallow blocks (blocks 0-19) causes only marginal degradation in AV synchronization. In contrast, disrupting the KV flow in middle or deep blocks (blocks 20-32 for audio and 25-47 for video) leads to a substantial drop in AV alignment, with the audio branch exhibiting a relative decrease of up to $\Delta 0.59$. This indicates that *shallow blocks are primarily responsible for intra-modal generation, while middle-to-deep blocks handle cross-modal audio-video interaction and alignment (AV-Sync Zone).*

Backward Analysis. Within the AV-Sync Zone, gradients flowing through the cross-modal interaction paths are dominant, indicating that the RL signal is correctly concentrated on the blocks truly responsible for cross-modal alignment. Meanwhile, the magnitudes of these interaction gradients gradually decay toward shallower layers. However, we observe an anomalous gradient spike in the shallow layers (Fig. 2(c)) on the audio branch, implying that a large portion of the RL gradient is erroneously injected into layers dedicated to intra-modal generation. We argue that this gradient misalignment disrupts the intra-modal audio generation process, resulting in sub-optimal results.

4.3 V2A Cross-Attention as an Intrinsic Proxy for Critical Regions

In joint audio-video generation, the local visual quality of sound-emitting regions plays a decisive role in shaping the subjective perceptual experience. However, uniform updates overlook their unequal contributions to the overall quality. This motivates us to localize such critical regions and apply differentiated levels of importance and exploration to them. While directly incorporating an external detection module offers a straightforward solution, it is computationally expensive. Instead, we observe an intrinsic indicator by analyzing the attention maps of the cross-modal attention layers within the audio branch, as illustrated in Fig. 4. We observe that these attention maps effectively highlight the speaking subjects and their sound-emitting regions within video frames. This strong correlation suggests that the *V2A cross-attention map naturally serves as an intrinsic proxy for identifying critical regions in video frames.*

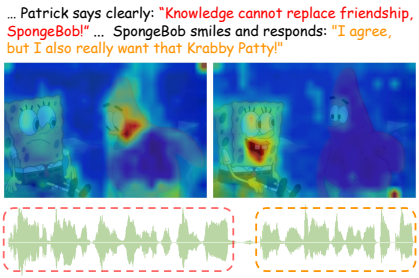

Figure 4: Visualization of V2A cross-attention maps.

5 Methodology

Motivated by the three issues and observations discussed above in Sec. 4, we propose **OmniNFT**, which performs fine-grained credit assignment at three corresponding levels: modality-wise advantage routing (Sec. 5.1), layer-wise gradient surgery (Sec. 5.2), and region-wise loss reweighting (Sec. 5.3). Fig. 5 illustrates the pipeline, and Alg. 1 summarizes the training procedure.

5.1 Modality-wise Advantage Routing

Reward-wise advantage computation. Instead of deriving a single advantage from all rewards, OmniNFT computes an independent advantage for each reward. During each sampling step, the model generates a group of G joint audio-video pairs $\{(x_v^{(i)}, x_a^{(i)})\}_{i=1}^{G}$ conditioned on prompt c. For each reward function $k \in \{v, a, av\}$, we evaluate the group to obtain raw scores $\{R_k^{(i)}\}_{i=1}^{G}$ and compute a reward-wise advantage $A_k^{(i)}$ following Eq. 6, which produces three decoupled advantage sets $\{A_v^{(i)}\}_{i=1}^{G}$, $\{A_a^{(i)}\}_{i=1}^{G}$, and $\{A_{av}^{(i)}\}_{i=1}^{G}$

Modality-decoupled advantage routing. With the reward-wise advantages, OmniNFT routes each advantage to the branch(es) whose output it evaluates. Specifically, the video advantage A_v captures visual quality, which is governed by the video branch. Analogously, the audio advantage A_a reflects audio fidelity determined by the audio branch. In contrast, the synchronization advantage A_{av} measures the synchronization alignment between the two modalities, and is thus *broadcast* as

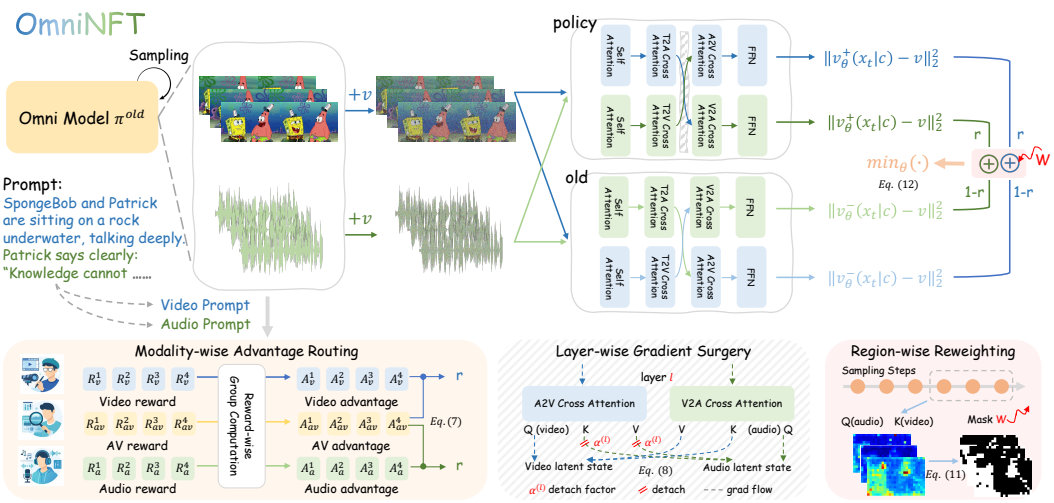

Figure 5: **Overview of OmniNFT.** Given paired video and audio prompts, the Omni model first generates joint audio-video samples. Building on these samples, OmniNFT performs three coordinated operations: (i) independent advantages derived from video, audio, and cross-modal rewards are dispatched to their corresponding branches (*Modality-wise Advantage Routing*); (ii) the audio-to-video cross-attention cached from the final sampling steps is converted into a critical-region mask that reweights the RL loss (*Region-wise Reweighting*); and (iii) during loss backward, the key-value gradients of A2V cross-attention in the shallow audio layers are partially detached, while all other gradient flows remain intact (*Layer-wise Gradient Surgery*).

shared supervision to both branches. The composite routed advantages are:

$$\tilde{A}_v^{(i)} = A_v^{(i)} + A_{av}^{(i)}, \qquad \tilde{A}_a^{(i)} = A_a^{(i)} + A_{av}^{(i)}. \tag{7}$$

This design enforces modality-specific supervision for uni-modal rewards while preserving shared cross-modal supervision, enabling more informative and less conflicting reward assignment.

5.2 Layer-wise Gradient Surgery

For each Transformer block l, the A2V cross-attention takes its query $Q_v^{(l)}$ from the video hidden states and its key-value pair $(K_a^{(l)}, V_a^{(l)})$ from the audio hidden states. We apply a layer-wise partial detach to these KV tensors:

$$\begin{aligned} \tilde{K}_a^{(l)} &= \alpha^{(l)} K_a^{(l)} + (1-\alpha^{(l)})\,\mathrm{sg}(K_a^{(l)}), \\ \tilde{V}_a^{(l)} &= \alpha^{(l)} V_a^{(l)} + (1-\alpha^{(l)})\,\mathrm{sg}(V_a^{(l)}), \end{aligned} \tag{8}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator and $\alpha^{(l)} \in [0,1]$ is the gradient retention factor. This operation leaves the forward computation unchanged while scaling the backward gradient through each modified KV path by $\alpha^{(l)}$. Based on the observed layer-wise functionality, we set:

$$\alpha^{(l)} = \begin{cases} \alpha_s, & l < L \quad \text{(shallow layers)}, \\ 1, & l \geq L \quad \text{(deep layers)}. \end{cases} \tag{9}$$

We use $L = 10$ and $\alpha_s = 0.1$ by default. Thus, the shallow V2A KV paths retain only 10% of their backward gradients, while the corresponding paths in deeper layers retain their gradients without attenuation. This design suppresses gradient propagation from the video stream through shallow audio KV connections while preserving the deeper cross-modal gradient pathways.

5.3 Region-wise Reweighting

Let $\mathrm{Attn}^{(l,u)} \in \mathbb{R}^{N_a \times N_v}$ denote the head-averaged V2A importance map at block l and sampling pass u, computed from scaled dot products of normalized Q_a and K_v, with softmax over K_v. We

Algorithm 1 OmniNFT: Modality-wise Omni Diffusion RL Fine-Tuning

Require: Pretrained policy $(\boldsymbol{v}_v^{\text{ref}}, \boldsymbol{v}_a^{\text{ref}})$, reward functions $\{R_v, R_a, R_{\text{av}}\}$, prompt dataset $\{\boldsymbol{c}\}$, gradient retention factor α_s, shallow boundary L, reweighting strength w, later denoising steps $\mathcal{T}$.
Initialize: $\boldsymbol{v}_m^{\text{old}} \leftarrow \boldsymbol{v}_m^{\text{ref}}$, $\boldsymbol{v}_{\theta_m} \leftarrow \boldsymbol{v}_m^{\text{ref}}$ for $m \in \{v, a\}$, data buffer $\mathcal{D} \leftarrow \emptyset$.

for each iteration i **do**
  for each sampled prompt $\boldsymbol{c}$ **do** *// Sampling Stage*
    Sample G joint outputs $\{(\boldsymbol{x}_{0,j}^v, \boldsymbol{x}_{0,j}^a)\}_{j=1}^G$ from $(\boldsymbol{v}_v^{\text{old}}, \boldsymbol{v}_a^{\text{old}})$.
    Evaluate rewards $\{R_v^{(j)}, R_a^{(j)}, R_{av}^{(j)}\}_{j=1}^G$ and cache V2A cross-attention maps $\{Attn^{(l,t)}\}$.
    Modality-wise advantages & routing (Sec. 5.1): compute $A_v^{(j)}, A_a^{(j)}, A_{av}^{(j)}$, then route: $\tilde{A}_v^{(j)} \leftarrow A_v^{(j)} + A_{\text{av}}^{(j)}, \quad \tilde{A}_a^{(j)} \leftarrow A_a^{(j)} + A_{\text{av}}^{(j)}$.
    Convert $\tilde{A}_m^{(j)}$ to optimality probability $r_m^{(j)}$ via Eq. 6 for $m \in \{v, a\}$.
    Region-wise weighting (Sec. 5.3): aggregate $\{Attn^{(j,l,t)}\}$ over $l \geq L, t \in \mathcal{T}$ to obtain $\{w^{(j)}\}$.
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{c, x_{0,j}^v, x_{0,j}^a, r_v^{(j)}, r_a^{(j)}, \{w^{(j)}\}\}$.
  end for
  for each mini batch $\{c, x_0^v, x_0^a, r_v, r_a, \{w^{(j)}\}\} \in \mathcal{D}$ **do** *// Training Stage*
    Layer-wise gradient surgery (Sec. 5.2): replace audio KV in A2V cross-attention block l with $\tilde{K}_a^{(l)} = \alpha^{(l)} K_a^{(l)} + (1 - \alpha^{(l)}) \cdot \text{sg}(K_a^{(l)})$, where $\alpha^{(l)} = \alpha_s$ if $l < L$ else 1, same for $\tilde{V}_a^{(l)}$.
    Compute region-weighted video loss and standard audio loss, then update θ via Eq. 12.
  end for
  Update $\theta^{\text{old}} \leftarrow \eta_i \theta^{\text{old}} + (1 - \eta_i)\theta$, and clear $\mathcal{D} \leftarrow \emptyset$. *// Online Update*
end for

Output: $(\boldsymbol{v}_{\theta_v}, \boldsymbol{v}_{\theta_a})$

aggregate collected blocks $\mathcal{B}$ and their recorded passes $\mathcal{C}_l$ within the selected denoising window:

$$s_i = \frac{1}{|\mathcal{B}|} \sum_{l \in \mathcal{B}} \frac{1}{|\mathcal{C}_l|} \sum_{u \in \mathcal{C}_l} \sum_{j=1}^{N_a} \text{Attn}_{j,i}^{(l,u)}, \quad i = 1, \ldots, N_v. \tag{10}$$

We standardize scores over spatial tokens within each latent video frame and compute:

$$\widehat{s}_i = \frac{s_i - \mu_{f(i)}}{\max(\delta_{f(i)}, \epsilon)}, \qquad w_i = 1 + \lambda_g \, \sigma(\lambda_g \widehat{s}_i), \tag{11}$$

where $f(i)$ indexes token i's latent frame, μ_f and δ_f are its spatial mean and sample standard deviation, $\epsilon = 10^{-8}$, and σ is the sigmoid function. The reweighting strength $\lambda_g \geq 0$ follows a linear warmup; weights are cached during sampling and fixed during optimization.

5.4 Overall Training Objective

Following the DiffusionNFT formulation (Sec. 3), we convert the routed advantages into optimality probabilities r_m for each branch $m \in \{v, a\}$. The region-wise weights w_i are incorporated into the video branch loss to concentrate optimization capacity on perceptually critical regions, while the audio branch loss remains unchanged. Based on Eq. 3, the total training objective becomes:

$$\mathcal{L}_{all}(\theta) = \underbrace{\frac{1}{\sum_i w_i} \sum_{i=1}^{N_v} w_i \cdot \mathcal{L}_{video}^{(i)}(\theta)}_{\text{region-weighted video loss}} + \mathcal{L}_{audio}(\theta). \tag{12}$$

6 Experiments

We first describe the experimental setup in Sec. 6.1, including reward models, evaluation metrics, and implementation details. We then present the main quantitative and qualitative results in Sec. 6.2, and conduct ablation studies to analyze the contribution of each component in Sec. 6.3. **We provide all the cases from the paper, as well as more scenarios, in the Supplementary material.**

Table 1: Main results on JavisBench (Liu et al., 2025c). Best results in green, second-best underlined. VQ: Visual Quality, AQ: Audio Quality. (↑: higher is better; ↓: lower is better).

Model	Size	AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony	
		VQ ↑	AQ ↑	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	AVHScore ↑	JavisScore ↑	DeSync ↓
– T2A+A2V											
TempoTkn (Yariv et al., 2024)	1.3B	–	–	0.084	–	0.205	–	0.139	0.122	0.103	1.532
TPoS (Jeong et al., 2023)	1.0B	–	–	0.201	–	0.229	–	0.124	0.129	0.095	1.493
– T2V+V2A											
ReWaS (Jeong et al., 2025)	0.6B	–	–	–	0.123	–	0.280	0.110	0.104	0.079	1.071
See&Hear (Xing et al., 2024)	0.4B	–	–	–	0.129	–	0.263	0.160	0.143	0.112	1.099
FoleyCrafter (Zhang et al., 2026)	1.2B	–	–	–	0.149	–	0.383	0.193	0.186	0.151	0.952
MMAudio (Cheng et al., 2025)	0.1B	–	–	–	0.160	–	0.407	0.198	0.182	0.150	0.849
– T2AV											
JavisDiT (Liu et al., 2025c)	3.1B	1.291	4.478	0.263	0.143	0.302	0.391	0.197	0.179	0.154	1.039
UniVerse-1 (Wang et al., 2025)	6.4B	1.357	4.839	0.272	0.111	0.309	0.245	0.104	0.098	0.077	0.929
JavisDiT++ (Liu et al., 2026a)	2.1B	1.462	5.049	0.282	0.164	0.316	0.424	0.198	0.184	0.159	0.832
LTX-2 (HaCohen et al., 2026)	19B	2.038	5.197	0.272	0.170	0.311	0.412	0.232	0.223	0.192	0.569
LTX-2+ GDPO (Liu et al., 2026b)	19B	3.209	5.523	0.265	0.184	0.308	0.428	0.233	0.223	0.185	0.412
LTX-2+OmniNFT	19B	3.326	5.715	0.261	0.189	0.310	0.445	0.262	0.257	0.220	0.269
Our RL Δ	–	+1.288	+0.518	-0.011	+0.019	-0.001	+0.033	+0.030	+0.034	+0.028	-0.300

6.1 Experimental Setup

Reward models. We use VideoAlign (Liu et al., 2025b) and HPSv3 (Ma et al., 2025) for video quality, Audiobox Aesthetics (Tjandra et al., 2025) for audio quality, CLAP (Wu et al., 2023b) for audio-text alignment, and DeSync (Iashin et al., 2024) for audio-visual synchronization.

Evaluation Metrics. We report results across four complementary dimensions defined in JavisBench: (i) *AV-Quality*: Visual quality (VQ) and audio quality (AQ) for uni-modal generation fidelity. Furthermore, we utilize VBench (Huang et al., 2024) as an additional benchmark to assess the quality of the generated videos. (ii) *Text-Consistency*: Text-Video and Text-Audio ImageBind (Girdhar et al., 2023) similarity (TV-IB, TA-IB), CLIP (Radford et al., 2021) score, and CLAP score for text-to-modal alignment; (iii) *AV-Consistency*: Audio-Video ImageBind similarity (AV-IB) and AVHScore for semantic coherence; (iv) *AV-synchronization*: JavisScore (Liu et al., 2025c) and DeSync for synchronization between audio and video.

6.2 Main Results and Analysis

Quantitative Analysis. Tab. 1 and Fig. 6 report results on JavisBench and VBench. Compared with LTX-2 (19B) and GDPO, OmniNFT achieves the best overall performance across perceptual quality, cross-modal consistency, and temporal synchronization. For perceptual quality, VQ improves from 2.038 to 3.326 (+63.2%) and AQ from 5.197 to 5.715 (+10.0%) over LTX-2, with notable VBench gains in imaging quality (+15.6%). For cross-modal consistency, TA-IB improves by 11.2% and AVHScore reaches 0.257, validating the effectiveness of modality-aware advantage routing. For temporal synchronization, OmniNFT reduces DeSync from 0.569 to 0.269 (-52.7%), substantially outperforming GDPO (0.412) Appendix A extends the comparison to MOVA (Team et al., 2026), while Appendix B reports lip-synchronization, speech-completeness, and human-preference results.

Setting	AV-Quality		Text-Consistency		AV-Consistency	
	VQ ↑	AQ ↑	CLIP ↑	CLAP ↑	AVHScore ↑	JavisScore ↑
(a) Gradient Surgery Layer Selection						
Shallow layers (default)	3.326	5.715	0.310	0.445	0.257	0.220
Deep layers	3.083	5.577	0.312	0.427	0.242	0.204
(b) Region-wise Reweighting Loss Weight λ						
$\lambda = 0.25$	3.150	5.495	0.308	0.429	0.249	0.212
$\lambda = 0.50$ (default)	3.326	5.715	0.310	0.445	0.257	0.220
$\lambda = 0.75$	2.977	5.714	0.310	0.438	0.260	0.223

Table 2: Ablation study on detach layers and coefficient λ.

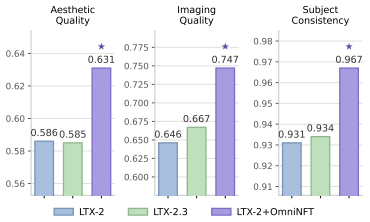

Figure 6: Vbench Results.

Qualitative Analysis. Fig. 7 illustrates visual quality in a cartoon scene, audio fidelity in a rooster scene, speech–lip synchronization in a close-up, and multi-speaker consistency in a confrontation scene. Additional examples appear in Appendix D; speech metrics and blind human preferences in Appendix B complement these qualitative comparisons.



Figure 7: Qualitative examples of joint audio-video generation by OmniNFT. The four cases illustrate improvements across different aspects: enhanced visual quality, improved audio fidelity, better speech-lip synchronization, and coherent multi-speaker scenes.

6.3 Ablation Studies

Effectiveness of Each Component. We validate the three key designs of OmniNFT by progressively incorporating them into the vanilla RL baseline, as shown in Tab. 3. Decoupling advantages across modalities alleviates reward inconsistency, bringing clear improvements in cross-modal consistency and synchrony. Shielding shallow audio layers from dominant video gradients further enhances audio fidelity and text-audio alignment, while focusing synchronization-critical regions provides fine-grained credit assignment that pushes AV-consistency and synchrony to the best results. These designs prove complementary, jointly resolving the optimization mismatch in joint audio-video RL fine-tuning. Notably, our designs introduce only a negligible overhead.

Hyperparameter Analysis. For gradient surgery layer L selection, detaching video-to-audio gradients at shallow layers ($L < 10$) consistently outperforms detaching at deep layers($L > 20$) in Tab. 2, which aligns with our observation. For the region-wise reweighting coefficient λ, a moderate value ($\lambda = 0.50$) achieves the best trade-off: smaller values under-emphasize critical regions, while larger values influence visual quality.

Table 3: Ablation results on Javisbench (↑: higher is better; ↓: lower is better).

	AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony		Efficiency
Model	VQ ↑	AQ ↑	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	AVHScore ↑	JavisScore ↑	DeSync ↓	Time ↓
LTX-2 (HaCohen et al., 2026)	2.038	5.197	0.272	0.170	0.311	0.412	0.232	0.223	0.192	0.569	–
LTX-2 + vanilla RL	3.209	5.523	0.265	0.184	0.308	0.428	0.233	0.223	0.185	0.412	23.9h
+ Modality-wise advantage routing	3.264	5.399	0.266	0.186	0.306	0.430	0.248	0.240	0.199	0.322	23.9h
+ Layer-wise gradient surgery	3.246	5.917	0.264	0.192	0.311	0.438	0.252	0.247	0.209	0.334	24.1h
+ Region-wise loss reweighting (Full)	3.326	5.715	0.261	0.189	0.310	0.445	0.262	0.257	0.220	0.269	24.1h

7 Conclusion

We present OmniNFT, a modality-aware online diffusion RL framework. Through three complementary innovations: modality-wise advantage routing, layer-wise gradient surgery, and region-wise loss reweighting, OmniNFT effectively resolves advantage inconsistency, gradient imbalance, and uniform credit assignment. Extensive experiments on JavisBench and VBench demonstrate our effectiveness.

AI USE STATEMENT

We used generative AI tools primarily to assist with translation and to improve the grammar, clarity, and readability of the manuscript. We reviewed the AI-assisted content to ensure accuracy and consistency with the underlying research. We take full responsibility for the final content of this work, including all text, claims, and results.

ETHICS STATEMENT

Data Usage. Our evaluation uses existing benchmarks, including JavisBench and VBench. Human preference results are reported in Appendix B.3. Any release of data or model artifacts will be subject to the applicable licenses, terms of use, and redistribution restrictions.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the following information:

1. The formulation and main components of OmniNFT are described in Sec. 5.
2. The experimental setup and evaluation metrics are presented in Sec. 6.1. Additional cross-backbone results are reported in Appendix A, while speech evaluation and human preference settings and results are provided in Appendix B.
3. The reported training configuration and computational budget are summarized in Appendix C.

We plan to release our code, model checkpoints, and training and evaluation configurations, subject to applicable licenses and redistribution restrictions. For third-party resources that cannot be redistributed, we will provide instructions for obtaining them.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 2024.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

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.

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

Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. Ace-step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.

Google DeepMind. Veo: A text-to-video generation system. Technical report, Google, 2024. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Accessed: 2025-09-24.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Xiaoxuan He, Siming Fu, Yuke Zhao, Wanli Li, Jian Yang, Dacheng Yin, Fengyun Rao, and Bo Zhang. Tempflow-grpo: When timing matters for grpo in flow models. *arXiv preprint arXiv:2508.04324*, 2025.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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

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

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7822–7832, 2023.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Yiming Cheng, Miles Yang, Zhao Zhong, and Liefeng Bo. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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. *arXiv preprint arXiv:2505.05470*, 2025a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025b.

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

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026a.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026b.

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

Yifu Luo, Xinhao Hu, Keyu Fan, Haoyuan Sun, Zeyu Chen, Bo Xia, Tiantian Zhang, Yongzhe Chang, and Xueqian Wang. Reinforcement learning meets masked generative models: Mask-grpo for text-to-image generation. *arXiv preprint arXiv:2510.13418*, 2025.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

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*, pp. 8748–8763. PmLR, 2021.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633, 2023a.

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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 2024.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, Yaning Pan, Hang Xu, and Feng Zhao. Maskfocus: Focusing policy optimization on critical steps for masked image generation. *arXiv preprint arXiv:2512.18766*, 2025a.

Guohui Zhang, Hu Yu, Xiaoxiao Ma, JingHao Zhang, Yaning Pan, Mingde Yao, Jie Xiao, Linjiang Huang, and Feng Zhao. Group critical-token policy optimization for autoregressive image generation. *arXiv preprint arXiv:2509.22485*, 2025b.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

A Additional Cross-Backbone Results

We evaluate OmniNFT on MOVA backbone to examine its effectiveness. The evaluation covers perceptual quality (VQ and AQ), text consistency (CLIP and CLAP), audio-video consistency (AVHScore), and temporal synchronization (JavisScore and DeSync). As shown in Table 4, OmniNFT outperforms GDPO on six of the seven reported metrics. In particular, DeSync decreases from 0.3380 to 0.2697, corresponding to a 20.21% relative reduction compared with GDPO and a 41.90% reduction compared with the base model.

Table 4: Additional results on MOVA. Best results in green. ↑ and ↓ indicate higher and lower is better, respectively.

	AV-Quality		Text-Consistency		AV-Consistency	AV-Synchrony	
Method	VQ ↑	AQ ↑	CLIP ↑	CLAP ↑	AVHScore ↑	JavisScore ↑	DeSync ↓
MOVA	3.1540	4.7357	0.3053	0.4385	0.2491	0.2112	0.4642
+ GDPO	4.0888	4.7734	0.3044	0.4520	0.2293	0.1977	0.3380
+ OmniNFT	3.9438	4.8030	0.3143	0.4650	0.2467	0.2117	0.2697

B Speech Evaluation and Human Preference

We complement the automatic audio-video evaluation with targeted assessments of lip synchronization, speech completeness, and human preference.

B.1 Lip Synchronization

We evaluate LTX-2 and its GDPO- and OmniNFT-tuned variants on 200 prompts featuring speakers from the MOVA Benchmark for Arena. We report Sync-C and Sync-D, for which higher and lower values indicate better lip synchronization, respectively. As shown in Table 5, OmniNFT achieves the highest Sync-C and the lowest Sync-D. Compared with GDPO, OmniNFT increases Sync-C by 0.063 and reduces Sync-D by 0.180.

Table 5: Lip-synchronization results on speaker prompts. Best results in green.

Method	Sync-C ↑	Sync-D ↓
LTX-2	5.034	9.022
+ GDPO	5.901	8.691
+ OmniNFT	5.964	8.511

B.2 Speech Completeness

We evaluate speech completeness on 200 speech prompts from the same benchmark using Qwen3-ASR. The default generation setting produces 121 frames at 24 FPS, corresponding to approximately five seconds and potentially limiting the completion of longer utterances. We report word error rate (WER), trailing-missing rate, and final-sentence completion rate.

Table 6 shows improvements in all three metrics. OmniNFT reduces WER from 0.1677 to 0.1128, a 32.74% relative reduction. The trailing-missing rate decreases by 6.91 percentage points, while the final-sentence completion rate increases by 6.54 percentage points. The latter two metrics measure different aspects of speech completeness and are reported separately.

Table 6: ASR-based speech evaluation on 200 speech prompts between OmniNFT and LTX-2.

Metric	LTX-2	OmniNFT	Δ
Word error rate ↓	0.1677	0.1128	−0.0549
Trailing-missing rate ↓	20.00%	13.09%	−6.91 pp
Final-sentence completion rate ↑	81.82%	88.36%	+6.54 pp

B.3 Blind Human A/B Preference Study

We conduct a blind A/B preference study using 100 prompts from the MOVA Benchmark for Arena. Each compared pair uses identical inference configurations, and the presentation order is balanced. The study assesses visual quality, audio quality, audio-video alignment, and overall preference. As shown in Table 7, OmniNFT receives higher preference rates than both the base model and GDPO across all reported dimensions.

Table 7: Blind A/B preference rates (%).

Dimension	OmniNFT / base model	OmniNFT / GDPO
Visual quality	67.41 / 32.59	57.18 / 42.82
Audio quality	66.06 / 33.94	54.64 / 45.36
Audio-video alignment	72.85 / 27.15	56.68 / 43.32
Overall	68.78 / 31.22	57.46 / 42.54

C Training implementation details

C.1 Training Data

We construct a training pool of 20,000 samples from VGGSound (Chen et al., 2020) using category-balanced sampling. Each selected sample is assigned to one of 13 semantic categories covering speech, music, animal sounds, environmental sounds, and other acoustic events. Table 8 summarizes the categories and their target sampling proportions.

The target distribution places greater emphasis on singing, human speech, and musical instruments while maintaining coverage of categories such as emergency alarms and infant or child sounds. Given the target sampling fraction p_c for category c, its target sample count is

$$N_c = p_c N_{\text{total}}, \qquad \sum_{c=1}^{13} p_c = 1,$$

where $N_{\text{total}} = 20{,}000$ is the size of the training pool. These proportions control the composition of the pool rather than the number of unique samples consumed during a particular training run.

C.2 Training and Inference Details

For the main LTX-2 experiments, we fine-tune the model using low-rank adaptation (LoRA) with Fully Sharded Data Parallel (FSDP). Training runs for 400 steps on 16 GPUs with a per-device batch size of 2 and a rollout group size of 8. Training videos are generated at a resolution of 512×768 (height $\times$ width), with 121 frames at 24 FPS. We use 20 sampling steps during training, with classifier-free guidance (CFG) scales of 1.5 for video and 3 for audio. For evaluation and inference, we use 40 denoising steps and increase the video and audio CFG scales to 3 and 7, respectively. Table 9 summarizes the training and inference configuration.

Computational overhead. Table 3 reports training times of 23.9 hours for vanilla RL and 24.1 hours for the full OmniNFT method under the listed configuration. This corresponds to an additional 0.2 hours, or approximately 0.84% of the vanilla RL training time.

Table 8: Target category distribution of the 20,000-sample training pool constructed from VGGSound.

Category	Proportion (%)	Description
Singing	20.0	Solo singing, choral singing, and humming.
Human Speech	20.0	Dialogue, lectures, narration, and reading aloud, excluding singing.
Musical Instrument	15.0	Instrumental performances, including piano, guitar, violin, drums, horns, and flutes.
Animal Vocalization	10.0	Sounds produced by animals, including dogs, cats, birds, frogs, and insects.
Other	10.0	Sounds outside the remaining categories.
Vehicle Transport	5.0	Transportation sounds from cars, trains, aircraft, motorcycles, and boats.
Household Daily	4.0	Cooking, eating, appliances, doors, footsteps, and typing.
Nature Environment	4.0	Wind, flowing water, rain, thunder, and other natural ambient sounds.
Electronic Music	3.5	Electronic beats, synthesizers, background music, and game sound effects.
Mechanical Industrial	3.0	Drills, saws, factory machinery, sewing machines, and related mechanical sounds.
Emergency Alarm	2.5	Emergency sirens, warning whistles, and alert buzzers.
Crowd Social	2.0	Cheering, applause, laughter, crowd chatter, and audience sounds.
Baby/Child	1.0	Infant crying or laughter, child speech, and play-related sounds.
Total	100.0	

Table 9: Training and inference configuration for the main LTX-2 experiments.

Configuration	Value
Training	
Fine-tuning strategy	LoRA
Distributed training framework	FSDP
Training pool size	20,000
Training steps	400
Number of GPUs	16
Batch size per device	2
Rollout group size	8
Video resolution ($H \times W$)	512×768
Frames per video	121
Frame rate	24 FPS
Sampling steps	20
Video CFG scale	1.5
Audio CFG scale	3
Evaluation / Inference	
Denoising steps	40
Video CFG scale	3
Audio CFG scale	7

D ADDITIONAL QUALITATIVE COMPARISONS

Figures 8–10 present additional qualitative comparisons between the base model and OmniNFT.

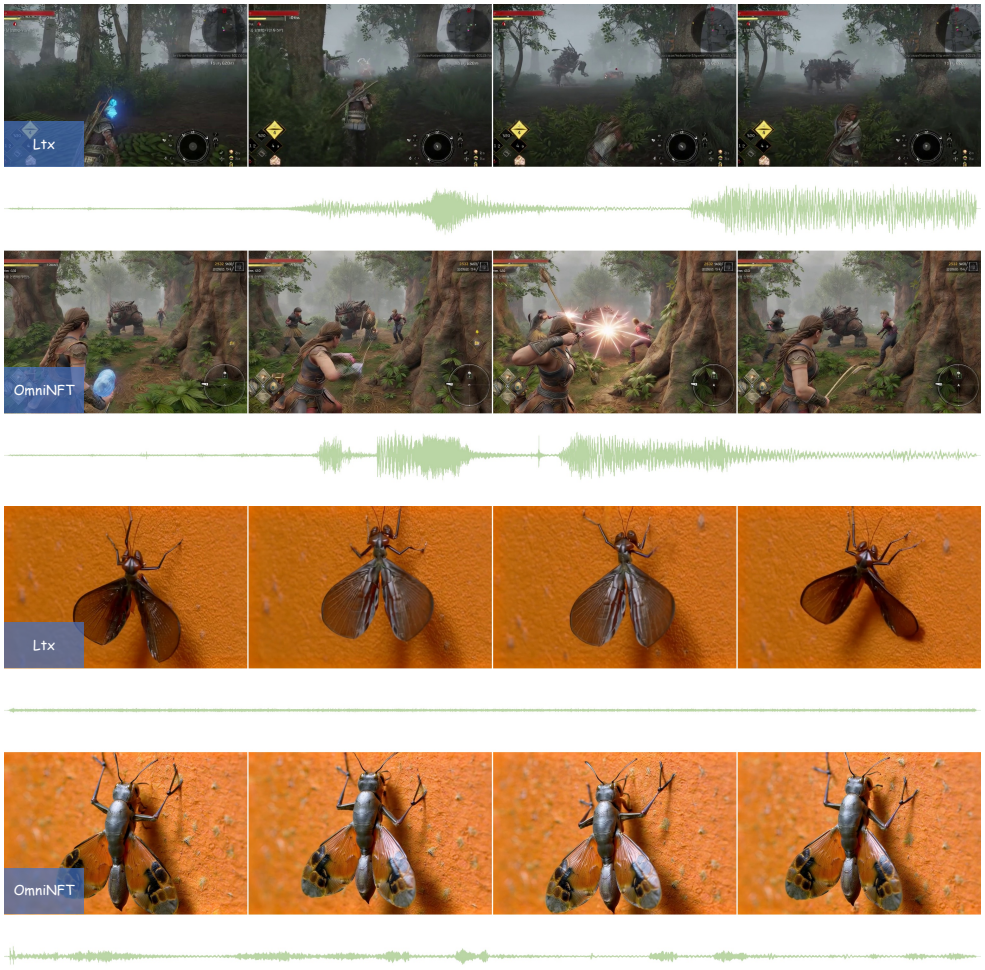

Figure 8: Additional qualitative comparisons for a game scene and an insect close-up.

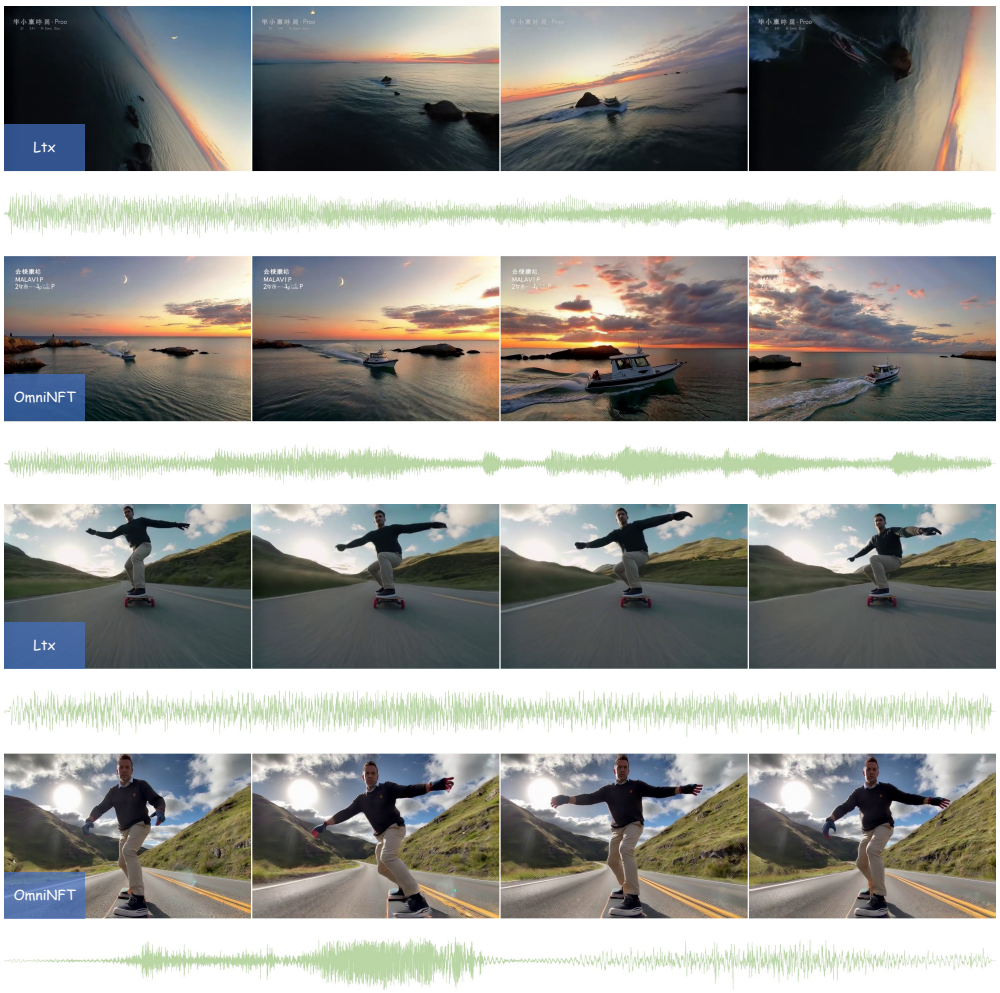

Figure 9: Additional qualitative comparisons for a boat scene at sunset and a skateboarding scene.

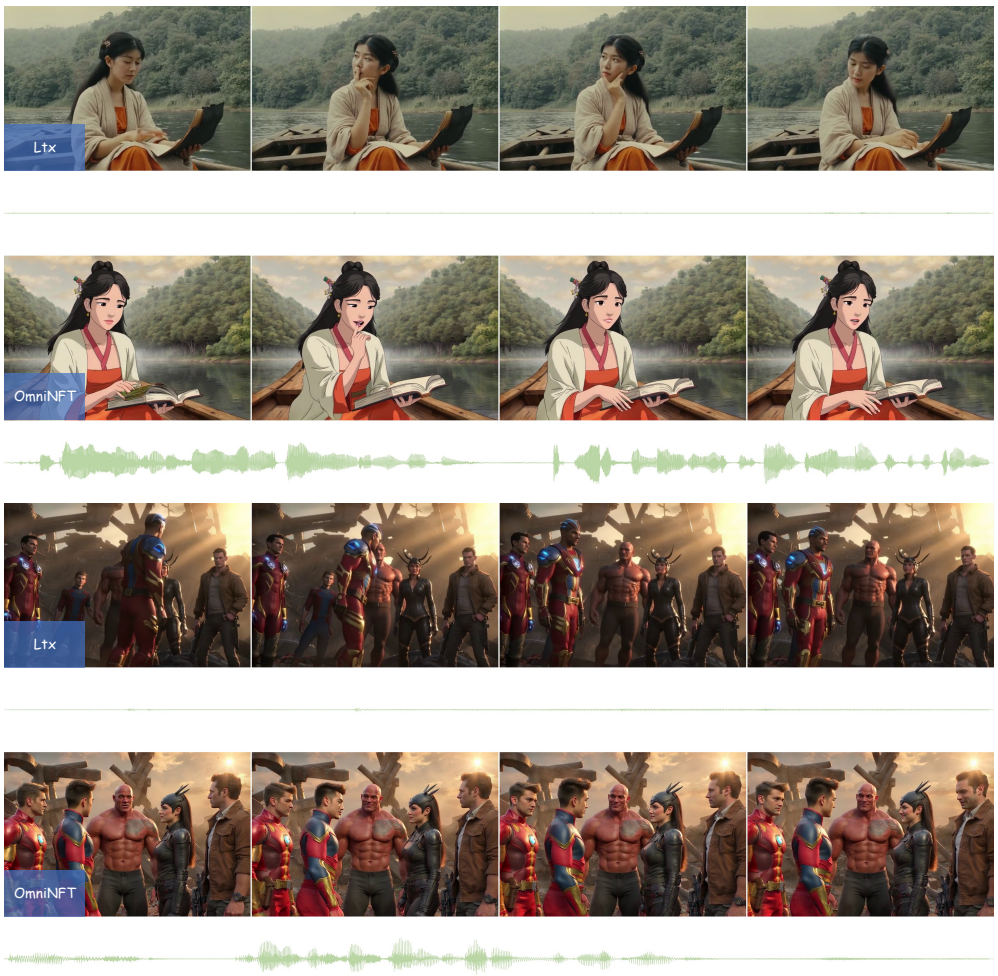

Figure 10: Additional qualitative comparisons for a woman in a boat and a multi-character scene.