OmniForcing: Unleashing Real-time Joint Audio-Visual Generation

Yaofeng Su[1,2*†], Yuming Li[3*], Zeyue Xue[1,4], Jie Huang[1], Siming Fu[1], Haoran Li[1], Haoyang Huang[1], and Nan Duan[1†]

[1] Joy Future Academy, JD, Beijing, China
[2] Fudan University, Shanghai, China
[3] Peking University, Beijing, China
[4] The University of Hong Kong, Hong Kong, China

Abstract. Recent joint audio-visual diffusion models achieve remarkable generation quality but suffer from high latency due to their bidirectional attention dependencies, hindering real-time applications. We propose OmniForcing, the first framework to distill an offline, dual-stream bidirectional diffusion model into a high-fidelity streaming autoregressive generator. However, naively applying causal distillation to such dual-stream architectures triggers severe training instability, due to the extreme temporal asymmetry between modalities and the resulting token sparsity. We address the inherent information density gap by introducing an Asymmetric Block-Causal Alignment with a zero-truncation Global Prefix that prevents multi-modal synchronization drift. The gradient explosion caused by extreme audio token sparsity during the causal shift is further resolved through an Audio Sink Token mechanism equipped with an Identity RoPE constraint. Finally, a Joint Self-Forcing Distillation paradigm enables the model to dynamically self-correct cumulative cross-modal errors from exposure bias during long rollouts. Empowered by a modality-independent rolling KV-cache inference scheme, OmniForcing achieves state-of-the-art streaming generation at ∼25 FPS on a single GPU, maintaining multi-modal synchronization and visual quality on par with the bidirectional teacher. **Project Page:** https://omniforcing.com.

Keywords: Streaming Audio-Visual Generation · Diffusion Distillation · Autoregressive Video Synthesis

1 Introduction

The landscape of generative AI has been significantly advanced by Diffusion Transformers (DiTs) [32,43]. Building on this foundation, joint audio-visual models such as LTX-2 [8] and Veo 3 [48] have recently achieved notable progress, leveraging modality-specific VAEs [8, 20, 23, 35, 44] to map video and audio into continuous latent spaces and jointly model their temporal distribution. However, this capability comes at a substantial computational cost. These models

* Equal contribution.
† Correspondence: yaofengsu@outlook.com, nanduan.nlp@outlook.com

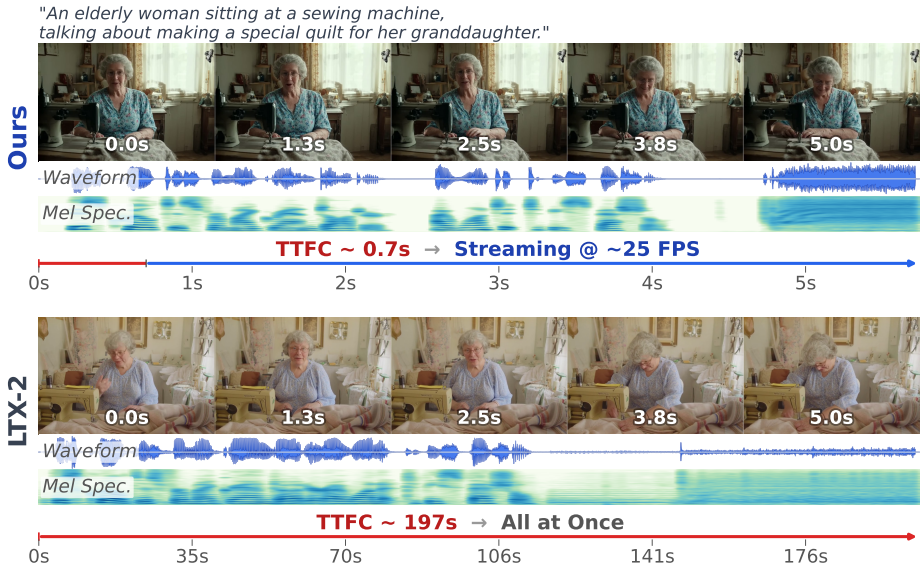

Fig. 1: OmniForcing breaks the latency barrier for joint audio-visual generation. Top: Our framework achieves real-time streaming at ∼25 FPS with an ultra-low Time-To-First-Chunk (TTFC) of ∼0.7s. Bottom: The bidirectional teacher (LTX-2) requires ∼197s to generate the sequence offline. OmniForcing maintains visual and acoustic fidelity on par with the teacher model.

rely heavily on bidirectional full-sequence attention, meaning the entire physical timeline must be processed simultaneously. Consequently, they suffer from a high Time-To-First-Chunk (TTFC) latency [55], limiting their deployment in interactive, real-time, or streaming applications (see Fig. 1 for a comparison).

To mitigate this latency bottleneck, previous efforts have diverged into two primary workarounds. The first line employs cascaded pipelines, generating video first and subsequently synthesizing audio [3, 30, 31, 46] (or vice versa) [14, 40]. However, this decoupled paradigm severs the joint distribution, limiting generation quality and fundamentally obstructing continuous streaming since the secondary audio modality cannot begin until the primary video has materialized sufficient context. Another line of work adapts video-only diffusion models into causal, autoregressive frameworks (e.g., CausVid [55], Self-Forcing [10]), but these methods remain confined to the visual domain. Directly extending them to dual-stream architectures is non-trivial, as the severe temporal asymmetry between modalities induces a critical information deficit for the sparser modality [15], leading to training instability rather than seamless stable alignment.

To address these challenges, we present **OmniForcing**, the first framework to successfully distill a heavy, bidirectional audio-visual foundation model (*i.e.*, LTX-2 [8]) into a high-fidelity streaming autoregressive generator. By dynamically interleaving the generation of audio and video chunks, OmniForcing enables

ultra-low latency streaming without sacrificing the holistic multi-modal distribution learned by the bidirectional teacher.

Our framework tackles the modality asymmetry through a carefully designed *Asymmetric Block-Causal Alignment*, and the full three-stage distillation pipeline is illustrated in Fig. 2. Firstly, to establish a persistent cross-modal anchor at the temporal origin, we introduce a zero-truncation *Global Prefix* that aligns the joint sequence at exact one-second boundaries (3 video frames to 25 audio frames) and exploits the native stride characteristics of the VAEs [8, 20]. Then, to stabilize the perilous cross-modal causal shift, we propose an unsupervised *Audio Attention Sink* mechanism, inspired by the attention sink phenomenon in language [50] and vision [4] models. Based on this design, we create a position-agnostic global memory buffer by assigning an *Identity RoPE* [39] constraint to these sink tokens, which mitigates the gradient explosions inherent in sparse causal audio attention. Furthermore, to combat the exposure bias amplified by cross-modal error accumulation during long rollouts, we employ a joint Self-Forcing [10] distillation strategy that enables the model to dynamically self-correct. Finally, we exploit the intra-layer decoupling between the 14B video and 5B audio streams [8] by introducing a *Modality-Independent Rolling KV-Cache* that reduces per-step context complexity to $\mathcal{O}(L)$ and enables concurrent execution of the two modality streams on a single GPU.

In summary, our main contributions are:

- We propose **OmniForcing**, a unified autoregressive framework that transforms offline, bidirectional joint audio-visual models into real-time streaming engines while preserving exact multi-modal temporal synchronization.
- We introduce a natural **Asymmetric Block-Causal Alignment** and the **Audio Sink Token** mechanism with Identity RoPE, providing a robust, position-agnostic solution to the Softmax collapse caused by multi-modal token density mismatch.
- We introduce a **Modality-Independent Rolling KV-Cache** with asymmetric parallel inference and a **Joint Self-Forcing Distillation** paradigm, which together mitigate exposure bias and reduce per-step context complexity to $\mathcal{O}(L)$, achieving state-of-the-art streaming generation at $\sim$25 FPS on a single GPU.

2 Related Work

Joint Audio-Visual and Video Foundation Models. The landscape of generative AI has been significantly advanced by large-scale Diffusion Transformers (DiTs) [32, 43]. For visual generation, foundation models such as Sora [28], Wan 2.1 [44], HunyuanVideo [49] and Kling [41] have demonstrated high visual fidelity, physical realism, and adherence to complex text prompts. Building upon these unimodal visual successes, the field has recently achieved a notable shift toward unified multimodal generation. Joint audio-visual foundation models like LTX-2 [8] and Veo 3 [48] have emerged as state-of-the-art systems capable of generating highly synchronized, high-fidelity audio and video in a single pass.

Notably, LTX-2 employs an asymmetric dual-stream architecture (a 14B video stream and a 5B audio stream) coupled through bidirectional cross-attention to deeply model the joint distribution of both modalities.

While these foundation models deliver remarkable semantic alignment and generation quality, they exhibit a critical limitation regarding deployment: they rely exclusively on *bidirectional full-sequence attention*. Because the generation of a single frame requires the model to simultaneously attend to the entire physical timeline, the computational complexity scales quadratically with sequence length. This results in a massive Time-To-First-Chunk (TTFC) latency, rendering these models ill-suited for supporting real-time, interactive, or streaming applications.

Audio-Visual Synthesis and Alignment. Prior to the emergence of joint foundation models, multi-modal generation heavily relied on cascaded or decoupled pipelines. These methods typically generate video first and subsequently synthesize the matching audio track using Foley sound generators (*e.g.*, FoleyGen [31], Diff-Foley [30], FoleyCrafter [56], MMAudio [3]), or build upon standalone audio foundation models like AudioLDM [23] and AudioGen [21]. Conversely, other works explore generating video driven by audio signals (A2V) [14, 40].

While computationally tractable, this decoupled paradigm inherently severs the joint temporal distribution. It struggles with fine-grained cross-modal synchronization and complex temporal reasoning, such as visual actions dynamically reacting to sudden acoustic events. Furthermore, this sequential video-to-audio (V2A) or audio-to-video (A2V) dependency fundamentally obstructs real-time streaming, a limitation OmniForcing avoids by streaming both modalities synchronously. Learnable context tokens have also been explored in audio-visual architectures for various purposes [2, 24], though none targets the causal distillation setting.

Diffusion Distillation for Efficiency. To break the latency barrier, various distillation methods compress multi-step diffusion sampling into one or a few evaluations. Distribution Matching Distillation (DMD) [53, 54] minimizes an approximate reverse KL divergence between student and teacher; Consistency Models [29,38] enforce self-consistency along ODE trajectories; and Adversarial Diffusion Distillation [37] leverages discriminator-based losses. These methodologies provide the computational foundation for real-time generation.

Autoregressive & Streaming Diffusion Models. Building upon efficient few-step distillation, recent pioneering works have transformed offline diffusion models into streaming architectures. Early explorations like StreamingT2V [9] and Pyramid Flow [16] introduced frame-wise and pyramid-based autoregressive diffusion, yet remained constrained by multi-step sampling overhead. CausVid [55] first established the core paradigm for streaming diffusion by distilling a bidirectional video teacher into a causal student using an asymmetric DMD pipeline, achieving $\sim$9.4 FPS streaming generation. Following this, Self-Forcing [10] identified and solved the critical *exposure bias* problem in autoregressive video generation by forcing the model to unroll its own KV-cache predictions during

training. This foundation has rapidly catalyzed a family of "forcing" variants tailored for diverse autoregressive bottlenecks, including Causal-Forcing [57] for stricter causal consistency and Rolling-Forcing [27] for minute-level long-context generation. Concurrently, block-wise or semi-autoregressive approaches such as PAR [47] and NBP [34] explore parallelized next-block prediction in the visual domain, but these are single-modality methods that are directly trained on data without leveraging a bidirectional teacher.

However, while these works represent a significant step forward for streaming generation, they operate exclusively on unimodal (video-only) architectures. Achieving real-time streaming for *joint audio-visual* generation remains an open, highly compelling, and unresolved problem. Furthermore, naively porting these causal distillation paradigms to a dual-stream multimodal architecture leads to severe training instability. Due to the severe frequency asymmetry between audio and video (e.g., 25 FPS vs. 3 FPS), imposing a causal mask creates extreme token sparsity and severe conditional distribution shifts, triggering Softmax collapse and gradient explosions. Therefore, formulating a stable, architecture-aware distillation pipeline tailored specifically for joint audio-visual streaming is of paramount importance. Our proposed OmniForcing naturally addresses this exact gap.

3 OmniForcing

3.1 Problem Formulation and The OmniForcing Pipeline

Given a text prompt c, our goal is real-time, streaming joint generation of a temporally aligned video $\mathbf{V}$ and audio $\mathbf{A}$, mapped into independent latent spaces via modality-specific VAEs [8, 20, 23, 35]. Following the distillation paradigm established by CausVid [55] and Self-Forcing [10], we restructure a pretrained bidirectional dual-stream transformer (LTX-2 [8]) into a **block-causal autoregressive** framework, factorizing the joint distribution over $K{+}1$ synchronized blocks $\{\mathcal{B}_0, \ldots, \mathcal{B}_K\}$, where K is the total number of physical seconds generated:

$$p(\mathbf{V}, \mathbf{A} \mid c) = p(\mathcal{B}_0 \mid c) \prod_{k=1}^{K} p(\mathcal{B}_k \mid \mathcal{B}_{<k}, c). \tag{1}$$

This retrofit faces three core challenges: (i) the extreme frequency asymmetry (3 FPS video vs. 25 FPS audio) hinders conventional causal masking; (ii) restricting the global bidirectional receptive field to sparse causal history triggers Softmax collapse and gradient explosions, an instability that is disproportionately severe for the audio stream, as diminishing the bidirectional context to extremely few tokens fundamentally undermines the modeling of continuous tokens [15]; (iii) exposure bias during long rollouts [10] is amplified into cross-modal desynchronization. OmniForcing addresses all three through an Asymmetric Block-Causal Masking design coupled with a three-stage distillation pipeline, transferring the teacher's high-fidelity joint distribution to an ultra-fast causal engine.

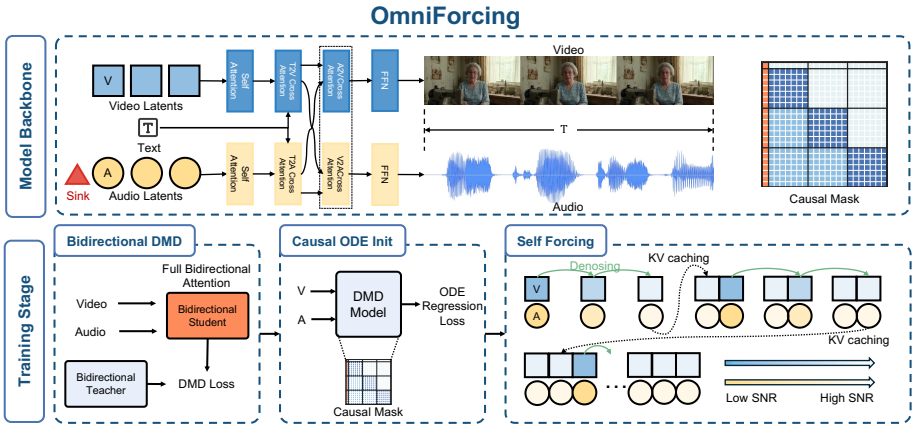

Fig. 2: The three-stage OmniForcing distillation pipeline. Stage I employs Distribution Matching Distillation (DMD) [53, 54] to adapt the model for few-step, fast denoising. Stage II utilizes causal ODE regression to adapt the network weights to the asymmetric block-causal mask. Stage III implements joint Self-Forcing [10] training by autoregressively unrolling the generation process to mitigate exposure bias.

3.2 Asymmetric Block-Causal Alignment and Mask Design

To achieve the block-level autoregressive factorization of the joint distribution, we must first bridge the inherent information density gap between the two modalities. In the physical world, video and audio exhibit starkly different spatiotemporal characteristics: video typically presents strong spatial redundancy and a low temporal evolution frequency, whereas audio is a dense high-frequency one-dimensional temporal signal. Consequently, when compressing into the latent space, the VAEs in multimodal generative models inevitably employ highly asymmetric temporal downsampling rates [8]. Specifically, within our LTX-2 [8] backbone, the video VAE outputs $f_v = 3$ latent frames per second, while the audio VAE outputs $f_a = 25$ frames per second.

Faced with this 25:3 non-integer frequency ratio, a strict frame-by-frame causal mask would lead to destructive feature truncation and temporal misalignment. To naturally resolve this conflict, we establish a physical-time-based **Macro-block Alignment**: given a one-second temporal window $\Delta T = 1\,\text{s}$, it perfectly encapsulates $\Delta N_v = f_v \times \Delta T = 3$ video latents and $\Delta N_a = f_a \times \Delta T = 25$ audio latents, without any fractional remainders.

The Global Prefix Token and the Mathematical Fit of VAE Strides. Notably, this block-level partitioning aligns closely with the causal convolutional stride characteristics of the underlying VAEs. In temporal compression, standard causal VAE architectures [8,20,44] typically apply an asymmetric treatment with a stride of 1 for the absolute first frame, while utilizing full-receptive-field strides for subsequent frames (e.g., a stride of 8 for video and 4 for audio). Therefore,

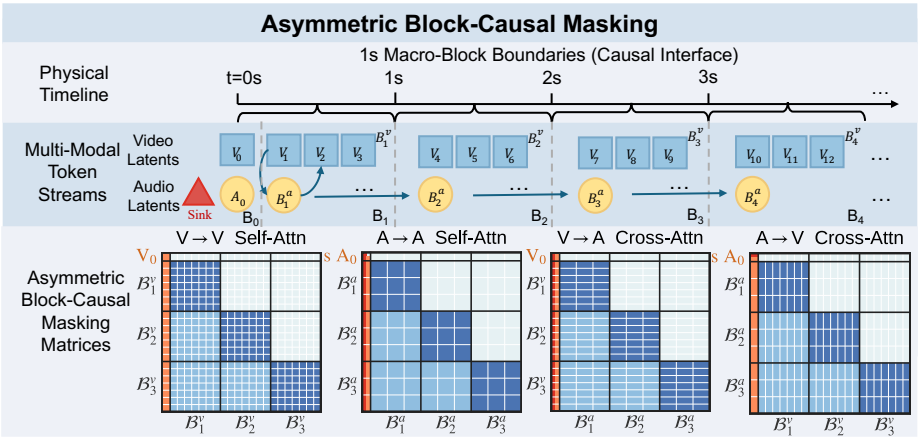

Fig. 3: Asymmetric Block-Causal Masking. The vertical axis denotes query tokens and the horizontal axis denotes key tokens. Modalities are synchronized via 1s macro-blocks. Each audio block (B^a) contains f_a latent frames (one token each; f_a=25), whereas each video block (B^v) contains f_v latent frames patchified into $f_v \cdot H_v W_v$ tokens (f_v=3, $H_v W_v$=384). Unmasked tokens include the Global Prefix (orange, V_0/A_0) and Audio Sink tokens (red, s). Blue regions denote allowed attention (bidirectional intra-block, strictly causal inter-block), while white regions mask future keys to prevent information leakage. Note that B_i^a and B_i^v denote the i-th audio and video macro-blocks, respectively, each spanning one physical second.

the length N of the entire latent sequence strictly follows the derived formulas:

$$N_v = 1 + K \cdot f_v, \quad N_a = 1 + K \cdot f_a, \tag{2}$$

where K is the total number of physical seconds generated (i.e., the total number of blocks). The constant term 1 in the formula corresponds to the initial latents $\mathbf{V}_0$ and $\mathbf{A}_0$ at $t \approx 0$ seconds.

Based on this derivation, the initial components $\mathbf{V}_0$ and $\mathbf{A}_0$ are naturally anchored at the origin in physical time, making it inherently impossible to squeeze them into the subsequent 1-second standard blocks $\mathcal{B}_k$. We transform this architectural inevitability into a design advantage by explicitly merging them into a **Global Prefix** block ($\mathcal{B}_0$). Within $\mathcal{B}_0$, the attention mechanism is unconditionally bidirectional. $\mathcal{B}_0$ functions similarly to a system prompt in large language models; it is immune to standard causal decay and remains globally visible to all future tokens. This not only ensures an elegant zero-truncation perfect alignment across the entire sequence but also provides a robust cross-modal semantic anchor for autoregressive long-sequence generation.

Natural Derivation of Four-Way Asymmetric Causal Masks. Having established the block-level alignment, we formalize the four-way attention masking strategy. Let $\tau(q)$ denote the physical block index of a given token q. For the video stream, each latent frame is spatially patchified into $H_v W_v$ tokens, where H_v, W_v are the spatial dimensions after VAE compression and patch embedding.

Each standard video block thus contains $|\mathcal{B}^v| = f_v \cdot H_v W_v$ tokens. For the audio stream, the mel-frequency bins are flattened into the channel dimension by the audio patchifier, so each latent frame maps to a single token; each standard audio block therefore contains $|\mathcal{B}^a| = f_a$ tokens. Accounting for the Global Prefix $\mathcal{B}_0$ (which holds $1 \times H_v W_v$ video tokens and 1 audio token), the block assignment for tokens beyond the prefix is:

$$\tau_v(q) = 1 + \left\lfloor \frac{q - H_v W_v}{f_v \cdot H_v W_v} \right\rfloor, \quad \tau_a(q) = 1 + \left\lfloor \frac{q-1}{f_a} \right\rfloor, \tag{3}$$

where tokens with indices below the prefix boundary ($q < H_v W_v$ for video, $q < 1$ for audio) belong to $\mathcal{B}_0$.

To enforce strict causality without future information leakage while allowing intra-block bidirectional flow, we define the binary attention mask $\mathbf{M} \in \{0, 1\}$ for query token q and key token k across all four attention pathways:

1. *Intra-modal Self-Attention:*

$$\mathbf{M}^{V \to V}_{q,k} = \mathbb{I}(\tau_v(k) \leq \tau_v(q)), \quad \mathbf{M}^{A \to A}_{q,k} = \mathbb{I}(\tau_a(k) \leq \tau_a(q)), \tag{4}$$

2. *Cross-modal Attention:*

$$\mathbf{M}^{V \to A}_{q,k} = \mathbb{I}(\tau_a(k) \leq \tau_v(q)), \quad \mathbf{M}^{A \to V}_{q,k} = \mathbb{I}(\tau_v(k) \leq \tau_a(q)), \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Because the Global Prefix tokens $\mathbf{V}_0$ and $\mathbf{A}_0$ are assigned to block $\tau = 0$, they inherently satisfy $\tau(k) \leq \tau(q)$ for all subsequent queries $q \geq 0$, mathematically guaranteeing their status as a globally visible, unmasked semantic anchor. This formulation ensures that despite the severe token density mismatch, the temporal receptive fields of both modalities expand synchronously at the physical block boundaries. The resulting four-way mask is visualized in Fig. 3.

3.3 Bridging the Gap: Causal Regression and Architectural Stabilizers

Inspired by recent video-only autoregressive distillation works (e.g., CausVid [55]), we design the first two stages of the pipeline. The core idea of these two stages is to smoothly decouple the few-step denoising capability from the causal generation paradigm and inject them into the model sequentially, paving the way for subsequent real-time joint inference.

Stage I: Bidirectional DMD. We first employ Distribution Matching Distillation (DMD) [53,54] to distill the original pretrained model into a bidirectional student model requiring very few steps. A jointly trained critic s_{gen} estimates the score of the student's output distribution, and the loss is a weighted sum of the video and audio DMD objectives: $\mathcal{L}_{\text{Bi-DMD}} = \lambda_v \mathcal{L}^v_{\text{DMD}} + \lambda_a \mathcal{L}^a_{\text{DMD}}$. While preserving the original global attention receptive field, this stage endows the model with strong few-step denoising capabilities, thereby providing a high-quality,

easily regressible teacher trajectory for the subsequent causal architectural migration.

Stage II: Causal ODE Regression. Next, we equip the model with the block-causal masks defined in Sec. 3.2. To adapt the weights to the causal masking without the complexity of full generation, we regress the ODE trajectories of the Stage I teacher v_ϕ. Let $\mathbf{x}^{(t)} = [\mathbf{V}^{(t)}, \mathbf{A}^{(t)}]$ denote the joint noisy latent at flow-matching time $t \in [0, 1]$, and superscripts v, a denote the video and audio velocity predictions, respectively:

$$\mathcal{L}_{\text{ODE}} = \mathbb{E}_{t,\mathbf{x}^{(t)}} \left[\lambda_v \| v_\theta^v(\mathbf{x}^{(t)}, c) - v_\phi^v(\mathbf{x}^{(t)}, c) \|_2^2 + \lambda_a \| v_\theta^a(\mathbf{x}^{(t)}, c) - v_\phi^a(\mathbf{x}^{(t)}, c) \|_2^2 \right]. \tag{6}$$

This stage aims to correct the causal maladaptation of the model weights, teaching the model to perform effective denoising predictions by observing only the causal history.

Conditional Distribution Shift and Gradient Explosion Crisis. Crucially, directly applying causal masks to the dual-stream model leads to a catastrophic collapse of Stage II training. The root cause lies in the severe *conditional distribution shift* when transforming bidirectional pretrained knowledge into the causal domain. The conditional distribution abruptly shifts from a globally-informed posterior to a truncated causal one:

$$\underbrace{p(\mathbf{x}_i \mid \mathbf{x}_{1:N}, c)}_{\text{Bidirectional (pretrained)}} \longrightarrow \underbrace{p(\mathbf{x}_i \mid \mathbf{x}_{1:i}, c)}_{\text{Causal (target)}}. \tag{7}$$

This information deficit is asymmetric across modalities. Because video utilizes spatiotemporal patch partitioning, a single physical block still contains hundreds of tokens (e.g., $f_v \times H_v W_v = 3 \times 384$ for our configuration), possessing a relatively abundant local context. Audio, however, usually has far fewer tokens per block; in our setting it contains a mere $\Delta N_a = f_a = 25$ tokens per block.

This extreme token sparsity destabilizes the attention mechanism. For audio tokens in early blocks, the visible history length is exceedingly small—the first token of a new block can only attend to itself and a few preceding tokens. With such a minuscule normalization denominator, the Softmax distribution degenerates into a near-one-hot vector with entropy approaching zero. In this saturated regime, minor logit perturbations are sharply amplified through the exponential nonlinearity, causing gradient variance to surge explosively ($\|\nabla \mathcal{L}\| \to \infty$) and producing NaN losses under fp16/bf16 precision.

Architectural Stabilizer: Audio Sink Tokens with Identity RoPE. To address this instability at its root, we introduce an effective architectural stabilizer, inspired by the attention sink phenomenon observed in autoregressive language modeling [50] and vision models [4]. We prepend S learnable Sink Tokens to the front of the audio sequence and permanently anchor them within the global prefix ($\mathcal{B}_0$).

In a physical sense, they act as a soft global memory buffer; mathematically, they forcefully expand the attention denominator for early audio tokens from i to $i+S$. This operation successfully breaks the extremely sharp Softmax collapse

and restores attention entropy, thereby acting as a contextual buffer to absorb anomalous logit perturbations and quelling the gradient storm at its source.

Furthermore, to prevent standard Rotary Position Embeddings (RoPE) [39] from imparting spurious physical temporal biases to these abstract tokens, we specifically enforce an *Identity RoPE* constraint on them:

$$\cos(\theta_{\text{sink}}) = \mathbf{1}, \quad \sin(\theta_{\text{sink}}) = \mathbf{0}. \tag{8}$$

Under this constraint, the standard rotary transformation analytically degenerates into an identity mapping ($\text{RoPE}(\mathbf{x}) = \mathbf{x}$). This shields the Sink Tokens from any positional rotational interference, making them position-agnostic semantic anchors that help stabilize the model across the causal transition.

3.4 Joint Self-Forcing Distillation and Asymmetric Parallel Inference

Although Stage II successfully injects causal generation capabilities, autoregressive models are inherently subject to exposure bias during long-sequence rollouts: because the model must condition on its own imperfect past outputs (rather than ground-truth data) during inference, prediction errors rapidly accumulate.

Stage III: Causal DMD with Joint Self-Forcing. To mitigate exposure bias, we introduce a joint audio-visual Self-Forcing [10] paradigm. Rather than conditioning on ground-truth history, the model unrolls the sequence autoregressively during training. Let G_θ denote the causal student generator and R_ϕ the frozen bidirectional teacher. Let $\hat{\mathcal{B}}_k$ be the block generated by G_θ; its KV embeddings are computed without noise and appended to a rolling cache. The joint Self-Forcing DMD loss evaluates the generated trajectory against the frozen bidirectional teacher R_ϕ via the DMD score-matching objective [53]:

$$\mathcal{L}_{\text{SF}} = \sum_{k=1}^{K} \mathbb{E}_{\hat{\mathcal{B}}_{<k},\, t,\, \boldsymbol{\epsilon}} \left[\lambda_v \left\| \nabla_{\text{KL}}^{v,k} \right\|_2^2 + \lambda_a \left\| \nabla_{\text{KL}}^{a,k} \right\|_2^2 \right], \tag{9}$$

where $\hat{\mathbf{x}}_0^k = G_\theta(\mathbf{z}_k \mid \mathcal{C}_{<k}, c)$ is the student's block-k output conditioned on the self-rolled KV-cache $\mathcal{C}_{<k}$, $\hat{\mathbf{x}}_{(t)}^k = (1-t)\hat{\mathbf{x}}_0^k + t\boldsymbol{\epsilon}$ is the flow-matching interpolation, and $\nabla_{\text{KL}}^{m,k} = s_{\text{gen}}^m(\hat{\mathbf{x}}_{(t)}^k, t) - s_{\text{CFG}}^m(\hat{\mathbf{x}}_{(t)}^k, t)$ denotes the per-modality score difference between the critic s_{gen} and the CFG-weighted teacher score s_{CFG}, for $m \in \{v, a\}$. This coupled unrolling forces the video and audio streams to dynamically adapt to each other's prediction drifts, ensuring strict cross-modal synchrony.

Asymmetric Compute Allocation and Parallel Inference. Ultimately, OmniForcing achieves true real-time inference by combining the aforementioned autoregressive algorithm with a flexible parallel strategy. In our architecture, the 14B video branch and the 5B audio branch exhibit extreme compute asymmetry. Thanks to our decoupled design in the attention pathways, there is absolutely no data dependency between video self-attention and audio self-attention.

Within each transformer layer, the audio and video streams are computed through independent FFN sub-layers and only synchronize briefly at the cross-modal attention boundaries (A2V and V2A). This intra-layer decoupling, combined with a Modality-Independent Rolling KV Cache that reduces per-step context complexity to $\mathcal{O}(L)$—where L denotes the total number of latent frames within the cache window—enables real-time synchronized audio-visual generation at ∼25 FPS on a single GPU. Furthermore, since the two modality-specific FFN sub-layers within each layer share no data dependency, the architecture naturally lends itself to asymmetric tensor parallelism across devices—assigning more compute to the heavier video stream—providing a practical scaling path toward higher resolutions and longer sequences in multi-GPU deployments.

4 Experiments

In this section, we comprehensively evaluate OmniForcing against both bidirectional models and cascaded autoregressive baselines. Since our primary goal is to achieve high-fidelity streaming outputs, we focus our evaluation on visual generation quality, audio fidelity, and real-time inference efficiency.

4.1 Experimental Setup

Implementation Details. We build OmniForcing on top of LTX-2 [8] (14B video stream + 5B audio stream). Training uses 32 GPUs with bf16 precision, a global batch size of 32, and a learning rate of 2×10^{-5}. Stage I (Bidirectional DMD) and Stage III (Joint Self-Forcing DMD) each run for 2,000 steps; Stage II (Causal ODE Regression) runs for 3,000 steps. DMD training in Stages I and III employs backward simulation [53]. We use $S = 16$ Audio Sink Tokens with Identity RoPE from Stage II onward and set classifier-free guidance scales to $w_v = 3$, $w_a = 5$. Our training set combines Mixkit video clips with captions sourced from the Open-Sora-Plan project [22], along with audio captions from AudioCaps [19]. All captions are rewritten by Gemma 3 12B [17] to produce audio-visually coherent descriptions. Since our method distills from a pretrained teacher rather than training from scratch, this relatively compact dataset suffices for the distillation objectives.

Streaming VAE Decoding and Latency Protocol. All reported latency metrics (TTFC, FPS, Runtime) include full VAE decoding and vocoder processing. Benchmarks use a single NVIDIA H200, bf16, CUDA 12.8, PyTorch 2.9.1, with SDPA MemoryEfficient; timing uses `torch.cuda.synchronize()` barriers (3 warmup, 100 iterations; TTFC median = 0.71 s, FPS median = 24.8). Our inference is strictly causal end-to-end: for video, a *sliding-window decoder* uses only the current block and W preceding blocks; for audio, a *cumulative decoder* re-decodes from the start, yielding lossless output. As shown in Tab. 4, W=2 achieves 47.4 dB PSNR at 288 ms per block, while audio decoding completes in ≤ 14 ms. Both decoders run concurrently, confirming real-time streaming.

Evaluation Metrics. We evaluate on JavisBench [26], following its four-dimension protocol: *AV-Quality* (FVD [42], FAD [18]), *Text-Consistency* (TV-IB, TA-IB via ImageBind [7]; CLIP [33]; CLAP [6]), *AV-Consistency* (AV-IB, AVHScore), and *AV-Synchrony* (JavisScore [26], DeSync [12]). We additionally report Runtime, TTFC, and FPS to quantify streaming capability. Here, TTFC measures the wall-clock time to generate and decode the Global Prefix ($\mathcal{B}_0$) together with the first streaming block ($\mathcal{B}_1$); once these are emitted, subsequent blocks are decoded concurrently with generation, enabling uninterrupted streaming playback. For distillation fidelity analysis, we use VBench [11], which officially supports per-video evaluation on user-supplied content.

Baselines. We compare against methods across three paradigms (Tab. 1): *T2A+A2V* cascaded pipelines (TempoTokens [52], TPoS [14]), *T2V+V2A* cascaded pipelines (ReWaS [13], Seeing & Hearing [51], FoleyCrafter [56], MMAudio [3]), and *T2AV* joint models (MM-Diffusion [36], JavisDiT [25], UniVerse-1 [45], JavisDiT++ [26]). The bidirectional LTX-2 [8] serves as the teacher upper bound.

4.2 Main Results

Inference Efficiency. As shown in Tables 1 and 2, OmniForcing completes a 5-second 480p audio-visual clip in ∼5.7 s of wall-clock time, with a TTFC of ∼0.7 s and a sustained throughput of ∼25 FPS—a ∼35× speedup over the offline LTX-2 teacher (197 s) and the only method in our comparison that enables true streaming generation.

Audio-Visual Quality. On JavisBench (Tab. 1), OmniForcing attains an FVD of 137.2 and FAD of 5.7, surpassing all baselines except the bidirectional teacher (FVD 125.4, FAD 4.6) on FVD and remaining comparable to the strongest competitor JavisDiT++ (FVD 141.5, FAD 5.5) on FAD, demonstrating that the core distributional quality of both video and audio is well preserved through distillation. For text-consistency, OmniForcing achieves the best CLIP score (0.322) across all methods, surpassing both LTX-2 (0.318) and JavisDiT++ (0.316), while TV-IB (0.287) ranks second overall. The audio-text metrics TA-IB (0.162) and CLAP (0.401) are slightly below the teacher but remain competitive with JavisDiT++ (0.164 and 0.424). For cross-modal coherence, OmniForcing achieves an AV-IB of 0.269 and AVHScore of 0.254, both ranking second overall and substantially outperforming all non-teacher baselines (the nearest competitor scores 0.198 on AV-IB). Temporal synchronization is also well preserved, with a DeSync of 0.392 closely tracking the teacher's 0.384 and dramatically outperforming all cascaded and joint baselines (JavisDiT++: 0.832). We attribute the modest gaps relative to the teacher in consistency and synchrony to the restricted causal receptive field replacing the original bidirectional full-sequence attention, an inherent trade-off for achieving streaming capability. Overall, OmniForcing achieves a ∼35× runtime reduction while attaining the strongest or second-strongest scores across nearly all quality dimensions.

Further Results on VBench and Qualitative Analysis. Since comprehensive joint audio-visual benchmarks remain scarce, we further validate the visual

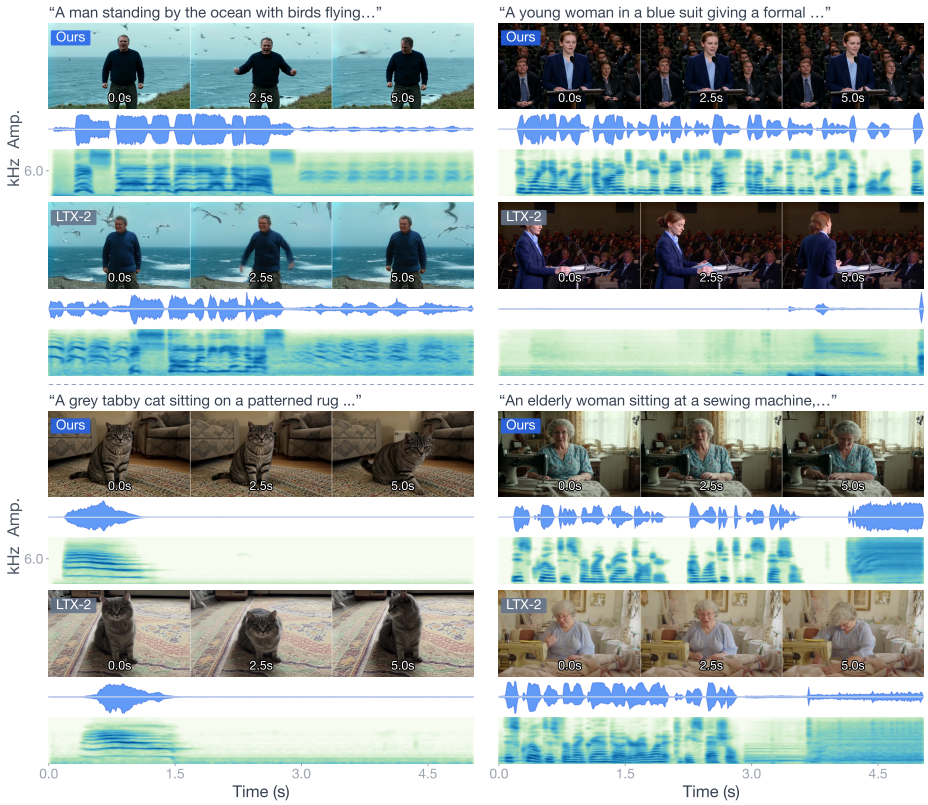

Fig. 4: Qualitative comparison across diverse scenes. Each example shows generated frames with synchronized waveforms and Mel spectrograms. OmniForcing produces voice layered with bird calls at a seaside scene (top-left), sustained speech at a podium presentation (top-right), a precisely timed cat meow (bottom-left), and blended narration with sewing machine sounds (bottom-right).

fidelity of our distillation using VBench [11] (Tab. 2). Under a controlled same-prompt comparison, OmniForcing slightly outperforms the teacher on per-frame quality metrics, including aesthetic quality (+0.026), imaging quality (+0.020), and subject consistency (+0.010), while temporal coherence (motion smoothness, temporal flickering) is likewise preserved. Although this may appear counterintuitive, it is consistent with prior distillation work [10, 53], which report that DMD-distilled students can surpass their teachers on per-sample quality metrics. On the audio side, qualitative results in Fig. 4 show that OmniForcing synthesizes layered voice-and-ambient sounds, sustained speech, synchronized transient events, and complex audio blends comparable to the teacher. These results together confirm that OmniForcing achieves a $\sim$35$\times$ latency reduction while maintaining the teacher's perceptual generation fidelity.

Table 1: Main results on JavisBench [25]. Best results in **bold**, second-best underlined. (↑: higher is better; ↓: lower is better). TTFC and FPS metrics are reported in Tab. 2.

		AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony		
Model	Size	FVD ↓	FAD ↓	TV-IB ↑	TA-IB ↑	CLIP ↑	CLAP ↑	AV-IB ↑	AVHScore ↑	JavisScore ↑	DeSync ↓	Runtime ↓
– T2A+A2V												
TempoTkn	1.3B	539.8	–	0.084	–	0.205	–	0.139	0.122	0.103	1.532	20s
TPoS	1.0B	839.7	–	0.201	–	0.229	–	0.124	0.129	0.095	1.493	19s
– T2V+V2A												
ReWaS	0.6B	–	9.4	–	0.123	–	0.280	0.110	0.104	0.079	1.071	17s
See&Hear	0.4B	–	7.6	–	0.129	–	0.263	0.160	0.143	0.112	1.099	25s
FoleyC	1.2B	–	9.1	–	0.149	–	0.383	0.193	0.186	0.151	0.952	16s
MMAudio	0.1B	–	6.1	–	0.160	–	0.407	0.198	0.182	0.150	0.849	15s
– T2AV												
MM-Diff	0.4B	2311.9	27.5	0.080	0.014	0.181	0.079	0.119	0.109	0.070	0.875	9s
JavisDiT	3.1B	204.1	7.2	0.263	0.143	0.302	0.391	0.197	0.179	0.154	1.039	30s
UniVerse-1	6.4B	194.2	8.7	0.272	0.111	0.309	0.245	0.104	0.098	0.077	0.929	13s
JavisDiT++	2.1B	141.5	5.5	0.282	0.164	0.316	0.424	0.198	0.184	0.159	0.832	10s
– T2AV & Autoregressive												
LTX-2	19B	**125.4**	**4.6**	**0.290**	**0.173**	0.318	**0.442**	**0.318**	**0.298**	**0.253**	**0.384**	197s
Ours	19B	137.2	5.7	0.287	0.162	**0.322**	0.401	0.269	0.254	0.208	0.392	**5.7s**

Table 2: Distillation fidelity: OmniForcing vs. the bidirectional LTX-2 teacher on the same prompt set. Visual quality is measured using VBench [11], which provides a modular toolkit that officially supports per-video scoring on user-supplied content. Best in **bold**. (↑: higher is better; ↓: lower is better).

	VBench Visual Quality					Streaming	
Model	Aesthetic Quality ↑	Imaging Quality ↑	Motion Smoothness ↑	Subject Consistency ↑	Temporal Flickering ↑	TTFC ↓	FPS ↑
LTX-2 [8]	0.569	0.574	0.993	0.945	0.988	197.0s	–
OmniForcing	**0.595**	**0.594**	**0.995**	**0.955**	**0.989**	**0.7s**	**25**

4.3 Ablation Studies

We isolate key components during Stage II, where the model first encounters the causal mask and instability is most pronounced. Results are summarized in Tab. 3.

Audio Attention Sink and Identity RoPE. We sweep the sink pool size $S \in \{1, 2, 4, 8, 16, 24\}$. Configurations with $S \geq 4$ converge stably with normal visual quality, while $S \leq 2$ triggers NaN gradients from Softmax collapse (Table 3), confirming that a sufficiently large sink pool is necessary to restore attention entropy. Replacing Identity RoPE with standard incremental positions at $S = 16$ yields convergence but elevated loss (0.402 vs. 0.081) and noisier outputs, validating that position-agnostic anchoring is essential.

Comparison with Alternative Stabilizers. QK-Norm [5] converges stably but its aggressive normalization dampens attention contrast, yielding higher loss (0.232). Tanh-Gated Attention [1] is more problematic: although no NaN occurs, the tanh saturation suppresses gradients so heavily that the loss plateaus at 1.258 and outputs degenerate into block artifacts—a signature of functionally destroyed attention. Neither alternative matches the stability–quality trade-off of our Audio Sink Tokens.

Streaming VAE Decoding Strategy. We ablate the sliding-window size W in Tab. 4. Quality saturates at W=2 (47.4 dB), matching W=4 at half the latency.

Table 3: Ablation on training stabilization strategies during Stage II. We report convergence status (with failure mode if applicable), maximum gradient norm, visual quality, and one-step denoising loss at $\sigma = 0.5$ averaged over the evaluation set after 3k steps. All sink variants use Identity RoPE unless noted.

Configuration	Convergence	$\|\nabla\|_{\max}$	Visualization	Loss ($\sigma=0.5$, 3k)	Remark
Sink $S=24$ + Id. RoPE	Stable	9.15	Normal	0.110	–
Sink $S=16$ + Id. RoPE	Stable	9.23	Normal	0.081	–
Sink $S=8$ + Id. RoPE	Stable	21.95	Normal	0.129	–
Sink $S=4$ + Id. RoPE	Stable	49.71	Normal	0.141	–
Sink $S=2$ + Id. RoPE	NaN	∞	Noise	–	Softmax collapse
Sink $S=1$ + Id. RoPE	NaN	∞	Noise	–	Softmax collapse
Sink $S=16$ + Incr. RoPE	Stable	11.21	Noisy	0.402	Positional bias
QK-Norm [5]	Stable	4.45	Normal	0.232	Damped contrast
Tanh-Gated Attn [1]	Plateau†	10.61	Block artifacts	1.258	Attention destroyed
No Stabilizer	NaN	∞	Noise	–	Softmax collapse

†No NaN observed, but loss does not decrease; the tanh gate functionally destroys attention patterns.

Table 4: Streaming VAE decode ablation. W: number of past latent blocks (no lookahead). Single H200, bf16, 3 warmup + 100 benchmark iterations. †Worst-case measured at 30 s.

Modality	Method	MAE↓	PSNR (dB)↑	Latency/Block
Video	Full sequence (ref.)	0	∞	–
Video	LTX-2 Causal	0.083	20.5	426 ms
Video	Slide $W=1$	0.065	19.8	201 ms
Video	Slide $W=2$ (Ours)	0.002	47.4	288 ms
Video	Slide $W=4$	0.002	47.4	444 ms
Audio	Full sequence (ref.)	0	∞	–
Audio	Cumul. (+vocoder, Ours)	0	∞	$\leq$14 ms†

LTX-2's built-in causal flag yields only 20.5 dB. Audio cumulative decoding is lossless at ≤ 14 ms.

5 Conclusion

We presented OmniForcing, the first framework to distill a bidirectional joint audio-visual diffusion model into a real-time streaming autoregressive generator. To overcome the temporal asymmetry and gradient instability inherent in causal multi-modal distillation, we introduced Asymmetric Block-Causal Alignment and Audio Sink Tokens with Identity RoPE, coupled with Joint Self-Forcing Distillation and a modality-independent rolling KV-cache. OmniForcing achieves ∼25 FPS streaming on a single GPU. We hope this work opens up new possibilities for deploying multi-modal foundation models in interactive and latency-sensitive scenarios.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Araujo, E., Rouditchenko, A., Gong, Y., Bhati, S., Thomas, S., Kingsbury, B., Karlinsky, L., Feris, R., Glass, J.R., Kuehne, H.: Cav-mae sync: Improving contrastive audio-visual mask autoencoders via fine-grained alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18794–18803 (2025)
3. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: 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)
4. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024)
5. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: International conference on machine learning. pp. 7480–7512. PMLR (2023)
6. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: Clap learning audio concepts from natural language supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
7. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: 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)
8. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
9. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
10. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
11. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., 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)
12. Iashin, V., Xie, W., Rahtu, E., Zisserman, A.: 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)
13. Jeong, Y., Kim, Y., Chun, S., Lee, J.: Read, watch and scream! sound generation from text and video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 17590–17598 (2025)
14. Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J.: The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceed-

ings of the IEEE/CVF international conference on computer vision. pp. 7822–7832 (2023)
15. Jia, D., Chen, Z., Chen, J., Du, C., Wu, J., Cong, J., Zhuang, X., Li, C., Wei, Z., Wang, Y., et al.: Ditar: Diffusion transformer autoregressive modeling for speech generation. In: International Conference on Machine Learning. pp. 27255–27270. PMLR (2025)
16. Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., MU, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: The Thirteenth International Conference on Learning Representations (2025)
17. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
18. Kilgour, K., Zuluaga, M., Roblek, D., Sharifi, M.: Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. arXiv preprint arXiv:1812.08466 (2018)
19. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 119–132 (2019)
20. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
21. Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., Adi, Y.: Audiogen: Textually guided audio generation. In: The Eleventh International Conference on Learning Representations (2023)
22. Lin, B., Ge, Y., Cheng, X., Li, Z., Zhu, B., Wang, S., He, X., Ye, Y., Yuan, S., Chen, L., et al.: Open-sora plan: Open-source large video generation model. arXiv preprint arXiv:2412.00131 (2024)
23. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: International Conference on Machine Learning. pp. 21450–21474. PMLR (2023)
24. Liu, K., Li, J., Sun, Y., Wu, S., gao, j., Zhang, D., Zhang, W., Jin, S., Yu, S., Zhan, G., Ji, J., Zhou, F., Zheng, L., Yan, S., Fei, H., Chua, T.S.: Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 142289–142324. Curran Associates, Inc. (2025)
25. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Jiang, R., Luo, J., Fei, H., et al.: Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
26. Liu, K., Zheng, Y., Wang, K., Wu, S., Zhang, R., Luo, J., Hatzinakos, D., Liu, Z., Fei, H., Chua, T.S.: Javisdit++: Unified modeling and optimization for joint audio-video generation. In: The Fourteenth International Conference on Learning Representations (2026)
27. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. In: The Fourteenth International Conference on Learning Representations (2026)
28. Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., He, L., Sun, L.: Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177 (2024)

29. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
30. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. Advances in Neural Information Processing Systems **36**, 48855–48876 (2023)
31. Mei, X., Nagaraja, V., Le Lan, G., Ni, Z., Chang, E., Shi, Y., Chandra, V.: Foleygen: Visually-guided audio generation. In: 2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2024)
32. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
34. Ren, S., Ma, S., Sun, X., Wei, F.: Next block prediction: Video generation via semi-autoregressive modeling. arXiv preprint arXiv:2502.07737 (2025)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N.J., Jin, Q., Guo, B.: Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10219–10228 (2023)
37. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
38. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Machine Learning. pp. 32211–32252. PMLR (2023)
39. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
40. Sung-Bin, K., Senocak, A., Ha, H., Owens, A., Oh, T.H.: Sound to visual scene generation by audio-to-visual latent alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2023)
41. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025)
42. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
45. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. arXiv preprint arXiv:2509.06155 (2025)
46. Wang, Y., Guo, W., Huang, R., Huang, J., Wang, Z., You, F., Li, R., Zhao, Z.: Frieren: Efficient video-to-audio generation network with rectified flow matching. Advances in neural information processing systems **37**, 128118–128138 (2024)

47. Wang, Y., Ren, S., Lin, Z., Han, Y., Guo, H., Yang, Z., Zou, D., Feng, J., Liu, X.: Parallelized autoregressive visual generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12955–12965 (2025)
48. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
49. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: Hunyuanvideo 1.5 technical report. arXiv preprint arXiv:2511.18870 (2025)
50. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: The Twelfth International Conference on Learning Representations (2024)
51. Xing, Y., He, Y., Tian, Z., Wang, X., Chen, Q.: 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)
52. Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., Adi, Y.: Diverse and aligned audio-to-video generation via text-to-video model adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6639–6647 (2024)
53. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
54. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6613–6623 (2024)
55. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
56. Zhang, Y., Gu, Y., Zeng, Y., Xing, Z., Wang, Y., Wu, Z., Liu, B., Chen, K.: Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. International Journal of Computer Vision **134**(1), 46 (2026)
57. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)