Taming Text-to-Sounding Video Generation via Advanced Modality Condition and Interaction

Kaisi Guan[*1,2], Xihua Wang[*1], Zhengfeng Lai[†2], Xin Cheng[1], Peng Zhang[2], Xiaojiang Liu[2], Ruihua Song[‡1], Meng Cao[2]

[1]Renmin University of China [2]Apple

guankaisi@ruc.edu.cn

Fig. 1: Examples of sounding videos generated by our BridgeDiT model, showcasing high quality, temporal synchronization, and text alignment. Our method generates high-fidelity video frames and detailed audio spectrograms that remain faithful to the given text prompts. Critically, as highlighted in the dashed boxes, the generated audio and video are precisely synchronized, demonstrating strong temporal coherence between visual events and their corresponding sounds.

Abstract. This study focuses on Text-to-Sounding-Video (T2SV) generation, which aims to generate a video with synchronized audio from text, with both modalities aligned to the text conditions. Despite progress in joint audio-video training, two critical challenges remain: (1) text conditioning is a bottleneck—shared captions ($T_V = T_A$) trigger modal interference, while a gap persists between dense training captions and concise inference user prompts, and (2) the optimal fusion mechanism for cross-modal feature interaction remains unclear. To address the first challenge, we first propose the Cross-Referential Rewriter (CRR) caption framework, a dual-agent pipeline where a Semantic Checker extracts grounded Semantic Anchors and a Cross-Modal Rewriter generates disentangled caption pairs (T_V and T_A), eliminating modal interference and

[1] *Equal contribution. Work done during an internship at Apple.

[2] †Project lead.

[3] ‡Corresponding author.

bridging the training-inference gap via prompt expansion. For the second, we introduce BridgeDiT, a dual-tower diffusion transformer that employs Dual Cross-Attention (DCA) as a bidirectional bridge between video and audio streams, which we show through systematic comparison to be the optimal fusion strategy for the dual-tower paradigm. Extensive experiments on three benchmarks, supported by human evaluations, demonstrate state-of-the-art results on most metrics. Comprehensive ablation studies further validate each component and offer key insights for future T2SV systems. The codes and models are available at https://bridgedit-t2sv.github.io/.

1 Introduction

Human perception is inherently multi-sensory, with vision and sound tightly coupled. Generating videos with synchronized audio from text (Text-to-Sounding-Video, T2SV) represents a crucial step toward world-modeling. Recent years have witnessed rapid progress in Text-to-Video (T2V) [2, 3, 48, 54, 63] and Text-to-Audio (T2A) [16, 22, 32, 50] generation. With these unimodal capabilities becoming increasingly mature, the community naturally shifts attention to the more challenging task of T2SV [23, 33, 37, 46, 55].

Prior strategies for T2SV suffer from critical limitations. Generating video and audio independently fails to achieve temporal synchronization. Pipelined methods (e.g., T→V→A or T→A→V) attempt to address this, but suffer from error accumulation—the second-stage model (Video-to-Audio [9, 10, 50, 53] or Audio-to-Video [24, 59, 60]), trained only on ground-truth data, cannot correct first-stage errors and often amplifies them. To overcome this, research has shifted toward joint generation, where both modalities are synthesized simultaneously. The single-tower paradigm [43, 45, 46, 51, 62] learns the joint distribution from scratch but is data-intensive and difficult to optimize. The dual-tower architecture [23, 33, 37, 52, 55] has thus emerged as the dominant approach, leveraging pretrained T2V and T2A backbones connected by a lightweight interaction module. Despite its promise, two fundamental challenges remain:

C1. The Conditioning Problem: Previous dual-tower methods [33, 37, 55, 62] employ a shared caption ($T_V = T_A$) for backbones pretrained on distinct modalities, causing semantic interference. Moreover, a significant training-inference gap exists: models are trained on dense captions, but users provide concise prompts at inference, creating a distribution shift that degrades performance.

C2. The Interaction Problem: The interaction module exchanges information between the two towers, but its optimal design remains unsolved. The core challenge is enabling effective bidirectional feature exchange to achieve both semantic and temporal synchronization.

We address the conditioning problem with the Cross-Referential Rewriter (CRR) caption framework. CRR is a dual-agent pipeline: a Semantic Checker cross-references raw captions from video and audio LLMs to extract grounded Semantic Anchors—filtering hallucinations and semantic conflicts—and a Cross-Modal Rewriter generates dense, disentangled captions (T_V, T_A) strictly from

these anchors. At inference, the same pipeline expands concise user prompts into modality-specific descriptions matching the training distribution, bridging the training-inference gap without complex prompt engineering. For the interaction problem, we propose BridgeDiT, a dual-tower architecture with Dual Cross-Attention (DCA) as its interaction module. Through systematic comparison with Full-Attention, Additive Fusion, and unidirectional alternatives under identical conditions, we identify DCA as the optimal fusion strategy: it preserves each tower's feature space while enabling symmetric, bidirectional information exchange, achieving strong synchronization without compromising generation quality. Extensive experiments on three benchmarks demonstrate state-of-the-art performance, supported by human evaluations. Detailed ablation studies validate the necessity of each component—Semantic Anchors, prompt expansion, and bidirectional fusion—and provide practical insights for T2SV system design. In summary, our contributions are:

- The CRR caption framework, a dual-agent pipeline that resolves the conditioning problem through grounded Semantic Anchors, modality-pure caption generation, and training-inference gap bridging via prompt expansion.
- The BridgeDiT architecture with DCA fusion, identified through systematic comparison as the optimal interaction strategy for the dual-tower T2SV paradigm.
- Comprehensive experiments and analyses across three benchmarks that achieve state-of-the-art results and offer key insights into caption design and fusion architecture choices.

2 Related Works

Single-Modal Generation. Recent years have seen significant progress in text-conditional single-modal generation, including Text-to-Video (T2V) [3, 28, 30, 48, 49, 54, 63] and Text-to-Audio (T2A) [16, 22, 32, 34]. Both domains have evolved from UNets [42] to Diffusion Transformers (DiT) [39] and adopted efficient paradigms like flow matching [31]. While these models produce high-quality individual outputs, generating video and audio independently often results in poor semantic and temporal synchronization. Cascaded pipelines address this by conditioning one modality on the other, such as Video-to-Audio (V2A) [9, 10, 11, 50, 53, 57] or Audio-to-Video (A2V) [24, 29, 59]. However, these sequential methods suffer from error accumulation [33, 37]: artifacts from the first stage inevitably propagate and degrade the final output [8, 18], motivating the shift toward joint audio-video generation.

Audio-Video Joint Generation. Existing methods follow two paradigms. The single-tower approach learns the joint audio-video distribution from scratch [43, 45, 46, 51, 62], but requires vast paired datasets and has shown limited practical success. The dual-tower paradigm has thus emerged as a more practical alternative, leveraging pretrained T2V and T2A backbones and focusing training on a lightweight interaction module for cross-modal fusion. The design of

this module is critical. Current strategies include full attention for direct fusion [52], ControlNet-style [61] unidirectional conditioning (video→audio [33] or audio→video [55]), and specialized components such as the Prior Estimator in JavisDiT [37]. Our work adopts the dual-tower paradigm and contributes a systematic comparison of fusion strategies, identifying Dual Cross-Attention as the optimal mechanism for bidirectional audio-video interaction.

Text Condition in Audio-Video Joint Generation. Text prompts serve as the fundamental semantic anchor for synchronizing audio and video modalities. Current strategies fall into two categories. *Shared prompting* methods such as CoDi [46] employ a single, usually visual-centric description to condition both modalities, inherently lacking auditory details. *Unified prompting* methods concatenate visual and auditory descriptions into one input. However, pretrained video and audio backbones are optimized for modality-specific text; feeding a fused prompt creates semantic interference, where visual attributes (*e.g.*, color) act as noise for the audio model and auditory descriptors (*e.g.*, timbre) distract the video model. Recent works LTX-2 [19] and Ovi [38] adopt this unified strategy but primarily target speech and dialogue, where lip-speech coupling naturally favors joint text representations. In contrast, our work focuses on foley and sound-effect generation, where the interference problem is more pronounced. We propose a *disentangled prompting* paradigm that generates modality-pure text conditions (T_V and T_A) to ensure precise alignment without interference.

3 Method

3.1 Preliminary

Denoising-based Generative Models Denoising generative models learn a complex data distribution $p(\mathbf{x})$ by reversing a process that destroys data to a simple Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Diffusion models [20] approach this by training a network $\boldsymbol{\epsilon}_\theta$ to predict the noise ϵ added to a data sample $\mathbf{x}_0$ at timestep t:

$$\mathcal{L}_{\text{DDPM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)||^2\right]. \tag{1}$$

Flow Matching (FM) [31] models learn a velocity field v_θ that transports a noise sample $\mathbf{x}_0$ to a data sample $\mathbf{x}_1$ by approximating the target field $\mathbf{x}_1 - \mathbf{x}_0$. The training objective is:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[||v_\theta(t\mathbf{x}_0 + (1-t)\mathbf{x}_1, t) - (\mathbf{x}_1 - \mathbf{x}_0)||^2\right]. \tag{2}$$

Generation in both cases involves starting with a sampled noise and applying the learned network to iteratively denoise and obtain a clean data sample. More background is in Appendix A.

Problem Formulation For the T2SV task, we adopt the dual-tower paradigm. This approach is highly practical as it leverages the capabilities of pre-trained unimodal models, a video tower $\mathcal{G}_\theta^V$ and an audio tower $\mathcal{G}_\theta^A$. In this setup, the towers independently process their respective text captions, T_V and T_A, audio timestep t_A and video timestep t_V, noisy audio latent $\mathbf{x}_A(t_A)$ and noisy video latent $\mathbf{x}_V(t_V)$ while a trainable interaction module, $\mathcal{B}_\theta^{AV}$, facilitates cross-modal communication:

$$(\hat{\mathbf{a}}, \hat{\mathbf{v}}) = \mathcal{G}_{\text{model}}(T_A, T_V, \mathbf{x}_A(t_A), \mathbf{x}_V(t_V), t_A, t_V), \quad \mathcal{G}_{\text{model}} = \{\mathcal{G}_\theta^A, \mathcal{G}_\theta^V, \mathcal{B}_\theta^{AV}\}. \quad (3)$$

The final output consists of the predicted audio $\hat{\mathbf{a}}$ and video $\hat{\mathbf{v}}$ noise vector.

Training Objective We sum up the loss from the two towers:

$$\mathcal{L} = \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{video}}.$$

The audio tower follows a diffusion training setup using a v-prediction diffusion [44] loss objective. Given the continuous timestep $t_A \in [0, 1]$, the signal and noise scaling factors are $\alpha(t_A) = \cos(t_A\pi/2)$ and $\sigma(t_A) = \sin(t_A\pi/2)$. We denote $\mathbf{x}_A$ as the audio latent vector from the audio Variational AutoEncoder (VAE) [27] encoder. It predicts the target $\alpha(t_A)\boldsymbol{\epsilon}_A - \sigma(t_A)\mathbf{x}_A$ and for the noisy audio latent $\mathbf{x}_A(t_A) = \alpha(t_A)\mathbf{x}_A + \sigma(t_A)\boldsymbol{\epsilon}_A$:

$$\mathcal{L}_{\text{audio}} = \left\|\hat{\mathbf{a}} - (\alpha(t_A)\boldsymbol{\epsilon}_A - \sigma(t_A)\mathbf{x}_A)\right\|^2.$$

The video tower follows a flow matching [31] loss objective. The corresponding video timestep t_V is defined as $t_V = 1000 \cdot t_A$. $\mathbf{x}_V$ is the video latent vector. It predicts the target vector field $\boldsymbol{\epsilon}_V - \mathbf{x}_V$ and for the noisy video latent $\mathbf{x}_V(t_V) = (1 - t_V/1000)\mathbf{x}_V + (t_V/1000)\boldsymbol{\epsilon}_V$:

$$\mathcal{L}_{\text{video}} = \left\|\hat{\mathbf{v}} - (\boldsymbol{\epsilon}_V - \mathbf{x}_V)\right\|^2.$$

Here, $\boldsymbol{\epsilon}_A$ and $\boldsymbol{\epsilon}_V$ are Gaussian noise vectors sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The detailed inference process is further shown in Appendix B.3.

3.2 Cross-Referential Rewriter Caption Framework

In the dual-tower paradigm, each tower requires its own text condition. Prior methods feed a shared caption to both towers [33, 37, 46, 62], causing semantic interference. A straightforward fix is to use an LLM to generate separate captions, but without cross-modal verification, audio LLMs frequently hallucinate sounds that have no visual source. The core issue is not caption generation itself, but the lack of a grounding mechanism between the two modal captions.

We address this with the **Cross-Referential Rewriter (CRR)** caption framework, which decomposes caption generation into two agents connected by a structured bottleneck. First, a Semantic Checker($\mathcal{F}_{sc}$) distills the input into *Semantic Anchors* $\mathcal{A}$—a structured set of verified attributes including core entities, environments, visual actions, and their corresponding acoustic events (*e.g.*,

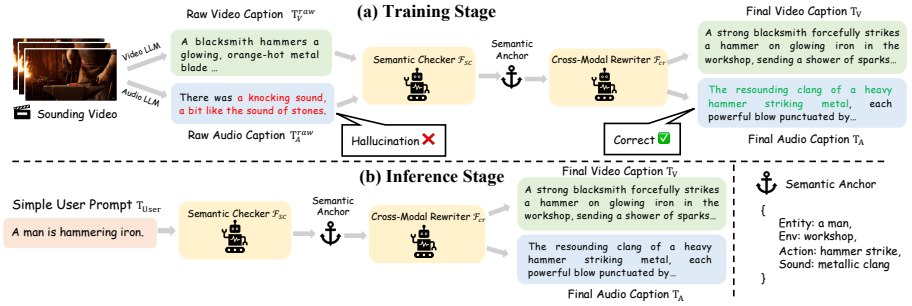

Fig. 2: Overview of the CRR caption framework. (a) Training: the Semantic Checker $\mathcal{F}_{sc}$ cross-references raw captions from Video and Audio LLMs to filter hallucinations (marked with red) and extract grounded Semantic Anchors $\mathcal{A}$ (structure shown on the right); the Rewriter $\mathcal{F}_{cr}$ then generates modality-pure captions (T_V, T_A). (b) Inference: the same pipeline expands a concise user prompt into dense, decoupled captions aligned with the training distribution.

`{Entity: worker, Env: workshop, Action: hammer strike, Sound: metallic clang}`). Then, a Cross-Modal Rewriter ($\mathcal{F}_{cr}$) generates dense, disentangled captions strictly from $\mathcal{A}$:

$$\mathcal{A} = \mathcal{F}_{sc}(\mathcal{I}), \quad T_V, T_A = \mathcal{F}_{cr}(\mathcal{A}). \tag{4}$$

Since $\mathcal{A}$ retains only grounded semantics and discards modality-specific noise, the Rewriter cannot introduce unverified events. This enforces modality purity by construction, not by relying on LLM instruction-following alone. The pipeline is uniform across training and inference—only the input $\mathcal{I}$ changes.

Training Stage: Given a sounding video, we first obtain independent raw descriptions from a Video LLM (Qwen2.5-VL [1]) and an Audio LLM (Qwen2-Audio [12]). While the Video LLM typically produces reliable visual descriptions, the Audio LLM is prone to hallucination—it often generates vague or incorrect sound descriptions due to the inherent ambiguity of audio signals. For example, given a worker hammering scene, the Audio LLM may output "there was a knocking sound, a bit like the sound of stones", failing to identify the actual sound source and introducing misleading acoustic content. Directly using such raw captions as training conditions would propagate these errors into the generative model. The Semantic Checker addresses this by cross-referencing both raw captions ($\mathcal{I} = (T_V^{raw}, T_A^{raw})$): it uses the visual description as a grounding reference to verify, correct, and filter the audio description, extracting only semantically consistent events into $\mathcal{A}$. The Rewriter then expands $\mathcal{A}$ into dense, modality-pure captions (T_V, T_A) that are both semantically aligned and free of cross-modal contamination.

Inference Stage: Users typically provide concise prompts (*e.g.*, "a worker is hammering iron") that are far shorter and less detailed than the dense training captions, creating a significant distribution gap. If fed directly to the model, such

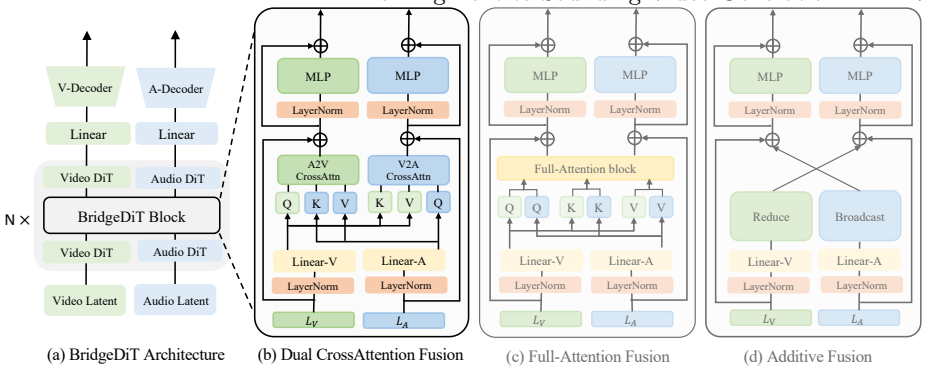

Fig. 3: Overview of the BridgeDiT Architecture. (a): The overall dual-tower architecture. Parallel video and audio DiT streams are connected by our proposed BridgeDiT Block at specific layers. *Right*: Details of fusion strategies within the block, showcasing our proposed Dual Cross-Attention (DCA) (b) alongside the Full-Attention (c) and Additive Fusion (d) baselines. The Unidirectional Cross-Attention baseline (omitted for brevity) shares the same structure as DCA but utilizes only a single direction (V2A or A2V) of the information flow.

brief inputs lead to degraded generation quality (validated in Table 4). Here $\mathcal{I} = T_{user}$, and the Semantic Checker applies the same verification logic but switches from cross-referencing to context inference: it deduces the implicit visual scenes and acoustic events entailed by the brief prompt (*e.g.*, "a worker striking iron" $\rightarrow$ $\mathcal{A}$=`{Entity: worker, Env: workshop, Action: hammer strike, Sound: metallic clang}`). The Rewriter then expands these anchors into detailed, disentangled T_V and T_A that match the density of training captions, bridging the distribution gap without requiring complex prompt engineering from users. Detailed agent instructions and system prompts for the CRR framework are provided in Appendix B.6.

3.3 Audio-Video Bridge Interaction Mechanism

A key challenge in the dual-tower paradigm is how to exchange information between the video and audio towers to achieve synchronization. This is nontrivial in our setting: both towers are pretrained on separate modalities and remain largely frozen, with features lying on two disjoint manifolds $\mathcal{M}_V$ and $\mathcal{M}_A$. This makes the fusion problem fundamentally different from models trained from scratch (*e.g.*, MMDiT [15]), where all modalities converge into a shared space.

BridgeDiT Architecture. As shown in Figure 3 (a), BridgeDiT inserts N lightweight interaction blocks between the parallel video and audio DiT streams. Each block performs cross-modal information exchange and returns updated features to both towers. We place the blocks at early-to-mid layers, where features balance spatial-temporal detail and semantic abstraction (see Appendix C).

Dual Cross-Attention Fusion (Ours). As shown in Figure 3 (b), our Dual Cross-Attention (DCA) mechanism updates the video latent L_V and audio latent

L_A through two symmetric cross-attention streams. Taking the Audio-to-Video (A2V) stream as an example, the video latent forms the query while the audio latent provides key and value:

$$Q_V = \text{Linear}_{Q_V}(\text{LN}(L_V)), \quad K_A = \text{Linear}_{K_A}(\text{LN}(L_A)), \quad V_A = \text{Linear}_{V_A}(\text{LN}(L_A)), \tag{5}$$

$$L'_V = \text{Attention}(Q_V, K_A, V_A) + L_V. \tag{6}$$

The V2A stream is perfectly symmetric: the audio latent serves as query, while the video latent provides key and value, yielding L'_A. While cross-attention is a well-established operation, we argue that DCA is the right choice for the pre-train backbone-based dual-tower setting for two reasons: (1) DCA preserves each tower's feature space—unlike Full-Attention, which concatenates both modalities and mixes intra- and inter-modal dependencies in one attention matrix, DCA keeps streams separate so each tower only receives targeted cross-modal cues; (2) this separation yields an easier optimization target—rather than merging $\mathcal{M}_V$ and $\mathcal{M}_A$ into a shared space with limited trainable parameters, DCA learns a lightweight bridge ($\mathcal{M}_V \leftrightarrow \mathcal{M}_A$) between two established manifolds. We validate this analysis in Section 4.5.

To justify our design, we compare Dual Cross-Attention Fusion against three baselines under identical conditions (Figure 3 (c,d)):
Full-Attention Fusion [52] concatenates both latents for joint self-attention, allowing all-to-all interaction but forcing a shared space:

$$L'_{\text{cat}} = \text{Attention}(\text{Cat}(Q_V, Q_A),\ \text{Cat}(K_V, K_A),\ \text{Cat}(V_V, V_A)) + \text{Cat}(L_V, L_A). \tag{7}$$

Additive Fusion [23] projects and adds cross-modal features to the residual stream—lightweight but non-selective:

$$L'_V = L_V + \text{Linear}_{A\to V}(\text{LN}(L_A)), \quad L'_A = L_A + \text{Linear}_{V\to A}(\text{LN}(L_V)). \tag{8}$$

Unidirectional Cross-Attention [33, 55] conditions one modality on the other in a ControlNet [61] style, restricting flow to one direction (V2A or A2V). Its formulation is identical to one half of DCA; we implement both variants.

4 Experiments

4.1 Experimental Setup

Implementation Details For the video backbone, we use Wan 2.1 (1.3B) [48] with a UMT5-XXL [13] text encoder, generating 81 frames at 15fps and 480p resolution. For the audio backbone, we employ Stable Diffusion Audio Open 1.0 [16] (SDA for short) with a T5-base text encoder [41], generating audio at

44.1kHz. The total generation length is 5.4 seconds. Our BridgeDiT architecture consists of 4 BridgeDiT Blocks uniformly inserted between the layers of both towers. For the CRR caption framework, both the Semantic Checker and Cross-Modal Rewriter are implemented using Qwen2.5-72B [58]; the raw captions are generated by Qwen2.5-VL-72B [1] (video) and Qwen2-Audio [12] (audio). We train our models separately for each dataset. More details are in Appendix B.

Datasets. We evaluate on three datasets: (1) AVSync15 [60], a subset of VGGSound [6] containing synchronized audio-video pairs across 15 categories, split into 1350 training and 150 test videos; (2) VGGSound-SS [5], a sound source localization dataset from VGGSound [6] with 5,158 videos across 220 classes, from which we randomly sample 150 for testing; and (3) Landscape [29], comprising 928 videos across 9 scenic categories. Since these datasets lack standard captions, we generate them using the CRR framework. All videos are standardized to 5.4 seconds via random cropping or padding.

Baselines. We compare against baselines spanning five T2SV generation paradigms. Since our work focuses on foley and sound-effect generation, we exclude speech-focused methods such as LTX-2 [19] and Ovi [38]. We distinguish two groups for fairness. *Retrained baselines* (Wan+SDA, SSVG [23], MTV [55]) are trained on each benchmark using our CRR captions, sharing identical data and text conditions with our method. *Pre-trained baselines* (MMAudio [9], SeeingHearing [57],ThinkSound [36],PrismAudio [35], JavisDiT [37], JointDiT [52], CoDi [46]) are evaluated with released checkpoints trained on substantially larger datasets; we feed them prompts in their original expected format to ensure optimal performance. The five paradigms are: (1) T→V‖T→A: Independent generation with interaction modules disabled (Wan+SDA). (2) T→V→A: Wan-1.3B generates video, then MMAudio [9] or SeeingHearing [57] synthesizes audio. (3) T→A→V: SDA [16] generates audio, then TPos [24] or TempoToken [59] produces video. (4) T→I→AV: Qwen-Image [56] generates an image for JointDiT [52] to jointly produce video and audio. (5) T→AV: Joint models including JavisDiT [37], SSVG [23], CoDi [46], and MTV [55]. Details are in Appendix B.2.

Automatic Evaluation Metrics We evaluate the T2SV task from three different perspectives: generation quality, text alignment, and audio-video synchronization. (1) Generation Quality. For video quality, we employ the Fréchet Video Distance (FVD) [47] and Kernel Video Distance (KVD) [47] with the Inflated 3D classifier [4]. For audio quality, we use the Fréchet Audio Distance [26] (FAD) and the Kullback-Leibler [53] (KL) divergence score. (2) Text Alignment. We evaluate video and audio text alignment separately. We use CLIPSIM [40]score to evaluate video-text alignment and CLAP [14] score to measure audio-text alignment. (3) Audio-Video Synchronization. We evaluate both semantic audio-video synchronization using the ImageBind score (VA-IB) [17] and temporal synchronization using the AV-Align score [59]. Recognizing the limitations of automatic metrics, we supplement these with human evaluations in Section 4.3.

Table 1: Automatic evaluation on the AVSync15 dataset. **Best** and second-best are highlighted.

Process	Method	Quality				Text Alignment		Synchronization	
		FVD↓	KVD↓	FAD↓	KL↓	CLIPSIM↑	CLAP↑	VA-IB↑	AV-Align↑
T → V ‖ T → A	Wan + SDA	828.33	22.56	11.90	3.17	28.12	30.78	26.22	0.205
T → V + V → A	Wan + MMAudio	828.33	22.56	7.98	1.40	28.12	34.64	33.50	0.243
	Wan + SeeingHearing	828.33	22.56	14.21	2.89	28.12	25.52	**35.87**	0.208
	Wan + ThinkSound	828.33	22.56	17.31	5.58	28.12	30.92	20.70	0.169
	Wan + PrismAudio	828.33	22.56	9.79	2.34	28.12	33.53	25.25	0.233
T → A + A → V	SDA + TPos	1975.22	92.95	11.90	3.17	18.64	30.78	25.45	0.176
	SDA + TempoToken	1516.53	42.30	11.90	3.17	20.56	30.78	17.63	0.215
T → I + I → VA	JointDiT	992.71	25.20	6.51	1.77	**29.94**	30.34	34.17	0.156
T → VA	JavisDiT	878.70	23.23	13.48	3.50	28.05	22.99	20.75	0.158
	SSVG	1028.78	31.44	14.35	4.35	25.78	23.67	22.45	0.126
	MTV	982.09	24.60	16.46	3.53	27.66	21.22	15.84	0.149
	CoDi	1387.14	39.24	16.56	5.24	24.96	17.94	10.79	0.081
	BridgeDiT (ours)	**765.74**	**21.33**	**5.34**	**1.30**	28.52	**35.95**	34.59	**0.275**

4.2 Main Results: Comparison with Baselines

We present results on three datasets in Table 1 and Table 2. Our BridgeDiT surpasses all baselines on most metrics, including video quality (FVD, KVD), audio quality (FAD, KL), audio-text alignment (CLAP), and temporal synchronization (AV-Align). Among the retrained baselines that share identical CRR captions and training data, BridgeDiT consistently outperforms SSVG and MTV, directly validating the architectural advantage of our DCA fusion mechanism. Compared with Wan+SDA—which is equivalent to our architecture with interaction modules removed—BridgeDiT shows clear improvements across all metrics, confirming that cross-modal interaction is essential for both generation quality and synchronization. BridgeDiT also outperforms all pipelined baselines, demonstrating that joint generation effectively mitigates the error accumulation inherent to sequential approaches. We note two minor exceptions. Our CLIPSIM (28.52) is slightly below JointDiT [52] (29.94), which we attribute to its stronger T2I backbone Qwen-Image [56]. Our VA-IB (34.59) is also surpassed by SeeingHearing (35.87), which directly uses ImageBind [17] as classifier guidance during inference—optimizing for this exact metric. As shown in Table 2, BridgeDiT achieves competitive results on VGGSound-SS and Land-

Table 2: Performance on VGGSound-SS and Landscape. AV denotes AV-Align metric here. **Best** and second-best are highlighted.

Method	VGGSound-SS			LandScape		
	FVD↓	FAD↓	AV-Align↑	FVD↓	FAD↓	AV-Align↑
T → V ‖ T → A						
Wan + SDA	737.96	8.05	0.238	700.18	5.80	0.177
T → V + V → A						
Wan + MMAudio	737.96	**5.39**	0.333	700.18	5.31	0.218
Wan + SeeingHearing	737.96	8.02	0.321	700.18	7.59	0.166
T → A + A → V						
SDA + TPos	1732.47	8.05	0.163	1837.42	5.80	0.143
SDA + TempoToken	1942.94	8.05	0.242	2089.05	5.80	0.206
T → I + I → VA						
JointDiT	866.59	7.19	0.125	937.09	6.35	0.075
T → VA						
JavisDiT	637.50	7.78	0.179	668.87	9.22	0.185
SSVG	1032.87	8.86	0.148	1186.39	8.97	0.143
CoDi	1203.68	8.38	0.113	1220.31	13.75	0.082
BridgeDiT (ours)	**615.28**	6.01	**0.362**	**628.07**	**4.78**	**0.258**

Table 4: Ablation study of the CRR caption framework on AVSync15.

Caption Strategy	FVD ↓	FAD ↓	CLIPSIM ↑	CLAP ↑	VA-IB ↑	AV-Align ↑
Shared text condition ($T_V = T_A$)						
Video Caption (T_V)	788.65	16.46	28.34	9.67	18.54	0.176
Audio-Video Caption (T_{AV})	1362.83	13.75	25.81	26.37	26.82	0.185
Disentangled text condition ($T_V \neq T_A$)						
Raw LLM Captions (w/o CRR)	924.36	19.42	26.22	8.43	17.84	0.161
Direct Rewrite (w/o Semantic Anchors)	787.32	15.74	28.34	27.34	28.82	0.224
CRR (ours)	**765.74**	**5.34**	**28.52**	**35.95**	**34.59**	**0.275**
CRR inference ablation						
w/o Prompt Expansion	1463.81	20.84	25.66	22.45	23.36	0.155

scape as well, confirming generalization across domains. However, automatic metrics do not always reflect perceptual quality—for instance, a method optimizing for a specific metric (*e.g.*, SeeingHearing for VA-IB) may score high without producing genuinely better outputs. To obtain a more reliable assessment, we conduct a comprehensive user study in Section 4.3.

4.3 User Study

To further validate our model with human preference, we conduct a user study on the AVSync15 test set with 150 samples. 20 evaluators rate the generated sounding videos on a 0–5 scale with 0.5-point increments across five criteria: Video Quality (VQ), Audio Quality (AQ), Text Alignment (TA), Synchronization (Sync), and Overall quality. As shown in Table 3, our BridgeDiT model is rated highest across all five dimensions, significantly outperforming all baselines, with the pipelined Wan+MMAudio ranking second. Notably, some baselines that achieve strong scores on specific automatic metrics (as in Table 1) are not preferred by human evaluators, reinforcing our observation that automatic metrics do not fully capture perceptual quality. This further highlights the importance of human evaluation as a complementary assessment for the T2SV task. Detailed annotation guidelines are provided in Appendix B.7.

Table 3: User study on the AVSync15 test set with 20 raters

Method	VQ ↑	AQ↑	TA↑	Sync↑	Overall↑
Wan+SDA	3.16	2.93	2.75	2.47	2.79
Wan+Seeing	3.16	2.83	2.91	2.54	2.85
Wan+MMAudio	3.16	3.18	3.07	2.77	3.06
SDA+TmToken	1.97	2.36	2.00	1.83	2.04
JointDiT	2.74	2.81	2.79	2.45	2.69
JavisDiT	2.47	2.39	2.36	2.14	2.36
BridgeDiT (ours)	**3.40**	**3.33**	**3.33**	**3.09**	**3.34**

4.4 Ablation Study on CRR Caption Framework

We ablate the CRR caption framework on AVSync15 by systematically varying the text conditioning strategy. Results are shown in Table 4. We organize the comparisons into three groups: shared text conditions, disentangled text conditions, and inference-stage ablation.

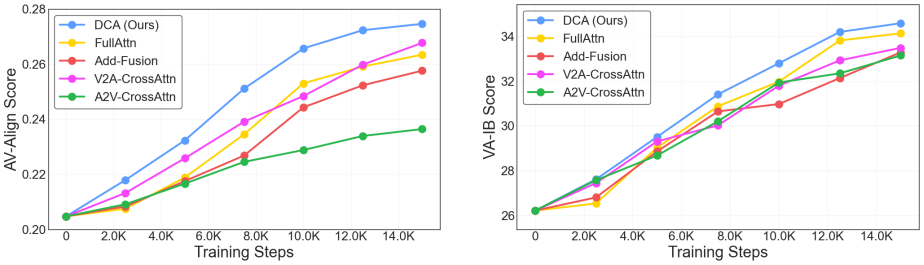

Fig. 4: Training dynamics of different fusion mechanisms on AV-Align and VA-IB scores. DCA achieves the highest synchronization throughout training.

Shared vs. Disentangled Conditioning. When both towers receive a shared caption, performance degrades regardless of caption content. Using only the video caption (T_V) starves the audio tower of acoustic details, resulting in poor audio-text alignment (CLAP: 9.67). The unified Audio-Video Caption (T_{AV}) retains some audio information but introduces semantic interference—visual attributes distract the audio model and vice versa—leading to severely degraded video quality (FVD: 1362.83). These results confirm that disentangled, modality-specific conditioning is essential for the dual-tower paradigm.

The Role of Semantic Anchors. Within the disentangled group, we compare three strategies of increasing sophistication. Raw LLM Captions directly use unverified outputs from the Video and Audio LLMs without any rewriting, yielding the worst disentangled performance (FAD: 19.42, AV-Align: 0.161) due to hallucinations and cross-modal conflicts. Direct Rewrite removes the Semantic Anchor bottleneck and lets a single LLM rewrite raw captions into modality-specific pairs in one step. This improves over raw captions, but still lags behind full CRR by a large margin, particularly in audio quality (FAD: 15.74 vs. 5.34) and synchronization (AV-Align: 0.224 vs. 0.275). The gap demonstrates that the structured Semantic Anchors are not merely a prompt engineering trick—they serve as a critical grounding mechanism that prevents the Rewriter from introducing unverified events, validating our core design argument in Section 3.2.

Inference-Stage Prompt Expansion. Removing prompt expansion at inference (*i.e.*, feeding concise user prompts directly) causes the most severe degradation (FVD: 1463.81, FAD: 20.84). This confirms that the distribution gap between brief user inputs and dense training captions is the single largest bottleneck for generation quality, and that CRR's prompt expansion is essential to bridge it.

4.5 Ablation Study on Fusion Mechanisms

We compare our DCA mechanism against four baselines under identical CRR captions and training conditions: Full-Attention [52], Additive Fusion [23], V2A-CrossAttn [33], and A2V-CrossAttn [55]. The comprehensive quantitative results of this evaluation are detailed in Table 5 while the corresponding training dynamics are visualized in Figure 4.

Table 5: Comprehensive evaluation on fusion mechanisms. We compare our DCA mechanism against various baselines on the AVSync15 dataset. Metrics include video quality (FVD), audio quality (FAD), text-alignment (CLIPSIM, CLAP), and audio-visual synchronization (VA-IB, AV-Align). **Bold** indicates the best performance.

Model	FVD ↓	FAD ↓	CLIPSIM ↑	CLAP ↑	VA-IB ↑	AV-Align ↑
No-Fusion	828.33	11.90	28.12	30.78	26.22	0.205
Full-Fusion	781.03	5.62	28.43	32.28	34.14	0.253
V2A-CrossAttn	813.02	6.21	28.24	35.85	34.20	0.268
A2V-CrossAttn	**746.37**	5.91	28.49	31.25	31.54	0.236
Additive-Fusion	772.23	5.72	28.20	34.77	28.34	0.258
DCA (Ours)	765.74	**5.34**	**28.52**	**35.95**	**34.59**	**0.275**

Results. The No-Fusion baseline performs worst on most metrics, confirming that cross-modal interaction is essential. Among all fusion strategies, DCA achieves the best overall performance: it ranks first on audio quality (FAD: 5.34), text alignment (CLIPSIM: 28.52, CLAP: 35.95), and both synchronization metrics (VA-IB: 34.59, AV-Align: 0.275). The A2V-CrossAttn variant achieves a slightly lower FVD (746.37), but at the cost of significantly weaker synchronization (AV-Align: 0.236), indicating that unidirectional flow improves one modality while neglecting the other. Figure 4 further shows that DCA consistently leads on both AV-Align and VA-IB throughout training, while other methods plateau or fluctuate at lower levels.

Why DCA Outperforms Full-Attention. Full-Attention (FA) is widely used for multimodal fusion in models trained from scratch (*e.g.*, MMDiT [15]), yet it underperforms DCA in our setting. We attribute this to the frozen-backbone constraint. In MMDiT, all parameters are trained jointly, allowing features from different modalities to gradually converge into a unified manifold $\mathcal{M}_{joint}$. In our framework, the video and audio backbones are independently pretrained and largely frozen, so their features reside on two disjoint manifolds $\mathcal{M}_V$ and $\mathcal{M}_A$. FA concatenates both latents into a single sequence and applies global self-attention, forcing intra-modal and inter-modal dependencies into one shared matrix. With limited trainable parameters, this effectively attempts to merge $\mathcal{M}_V$ and $\mathcal{M}_A$ into a shared space—a difficult optimization target that leads to the unstable synchronization curves observed in Figure 4. DCA, by contrast, keeps the two streams separate and only exchanges targeted cross-modal cues through dedicated attention heads. It learns a lightweight bridge ($\mathcal{M}_V \leftrightarrow \mathcal{M}_A$) between two established manifolds rather than attempting to unify them. This is a fundamentally easier task given the parameter budget, resulting in more stable training and stronger synchronization. This analysis also explains why the two unidirectional variants show asymmetric behavior: A2V-CrossAttn improves video quality (FVD: 746.37) by conditioning video on audio, but leaves the audio stream unrefined; V2A-CrossAttn shows the opposite pattern. DCA resolves this

by enabling bidirectional exchange, ensuring that both modalities benefit from cross-modal information simultaneously.

4.6 Case Studies

Figure 1 presents case studies that highlight the capabilities of our BridgeDiT model. Powered by the combination of our CRR caption framework and the DCA fusion mechanism, our model generates high-quality sound videos that are semantically synchronized, temporally synchronized, and highly aligned with the text conditions. The first case (the blacksmith) showcases precise temporal synchronization, as the visual impact of the hammer striking the iron aligns perfectly with the sharp "clang" event in the audio spectrogram. The second case (the saxophone player) demonstrates strong text alignment; the generated video accurately depicts key entities from the visual prompt, including the "saxophone" and "metal drum", while the audio faithfully synthesizes the complex soundscape described in the audio prompt. Notably, these two cases represent distinct challenges: the blacksmith involves sparse, impulsive sound events requiring precise temporal alignment, whereas the saxophone scene involves continuous, overlapping audio sources demanding accurate semantic disentanglement—our model handles both effectively. We provide more cases in the Appendix B.4.

5 Conclusion

In this work, we address two fundamental challenges in Text-to-Sounding-Video generation: suboptimal text conditioning and ineffective cross-modal interaction. For text conditioning, we propose the CRR caption framework, a dual-agent pipeline where a Semantic Checker extracts grounded Semantic Anchors and a Cross-Modal Rewriter produces disentangled, modality-pure captions, eliminating modal interference and bridging the training-inference gap. For cross-modal interaction, we propose BridgeDiT, connecting pretrained video and audio backbones via Dual Cross-Attention (DCA), which we show through systematic comparison to be the optimal fusion strategy for the dual-tower paradigm. Experiments on three benchmarks with human evaluations demonstrate state-of-the-art performance, and ablation studies validate the necessity of each component. We believe these findings provide a solid foundation for scaling T2SV systems to longer durations and more diverse acoustic scenarios.

Limitations. Our evaluation currently relies on relatively small-scale benchmarks (up to 5K videos), which may not fully capture the diversity of real-world acoustic environments. We also do not yet support speech or music generation: extending to these modalities is hindered by the scarcity of open-source pre-trained models and would require resource-intensive training from scratch. Additionally, exploring post-training refinement techniques, such as Reinforcement Learning from Human Feedback (RLHF), presents a promising pathway to further optimize audio-visual synchronization and align with human perception.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 62276268) and the Outstanding Innovative Talents Cultivation Funded Programs 2022 of Renmin University of China. Experiments of this work used the Bridges2 system at PSC and Delta system at NCSA through allocations CIS210014 and IRI120008P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

References

[1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)

[2] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), https://arxiv.org/abs/2311.15127

[3] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators

[4] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset (2018), https://arxiv.org/abs/1705.07750

[5] Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR (2021)

[6] Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2020)

[7] Chen, J., Zhao, Y., Yu, J., Chu, R., Chen, J., Yang, S., Wang, X., Pan, Y., Zhou, D., Ling, H., Liu, H., Yi, H., Zhang, H., Li, M., Chen, Y., Cai, H., Fidler, S., Luo, P., Han, S., Xie, E.: Sana-video: Efficient video generation with block linear diffusion transformer. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=mzAchylAtf

[8] Chen, L., Chen, H., Cai, Y., Li, S., Ye, Q., Wang, Y.: Detecting and mitigating insertion hallucination in video-to-audio generation (2025), https://arxiv.org/abs/2510.08078

[9] 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)

[10] Cheng, X., Wang, X., Wu, Y., Wang, Y., Song, R.: Lova: Long-form video-to-audio generation. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)

[11] Cheng, X., Wang, Y., Wang, X., Wu, Y., Guan, K., Chen, Y., Zhang, P., Liu, K., Cao, M., Song, R.: Vssflow: Unifying video-conditioned sound and speech generation via joint learning (2026), https://www.arxiv.org/abs/2509.24773

[12] Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J.: Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 (2024)

[13] Chung, H.W., Constant, N., Garcia, X., Roberts, A., Tay, Y., Narang, S., Firat, O.: Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. arXiv preprint arXiv:2304.09151 (2023)

[14] 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)

[15] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://arxiv.org/abs/2403.03206

[16] Evans, Z., Parker, J.D., Carr, C., Zukowski, Z., Taylor, J., Pons, J.: Stable audio open. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)

[17] 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: CVPR (2023)

[18] Guan, K., Lai, Z., Sun, Y., Zhang, P., Liu, W., Liu, K., Cao, M., Song, R.: Etva: Evaluation of text-to-video alignment via fine-grained question generation and answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21299–21309 (October 2025)

[19] 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)

[20] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020), https://arxiv.org/abs/2006.11239

[21] Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

[22] Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z., Zhao, Z.: Make-an-audio 2: Temporal-enhanced text-to-audio

generation (2023)

[23] Ishii, M., Hayakawa, A., Shibuya, T., Mitsufuji, Y.: A simple but strong baseline for sounding video generation: Effective adaptation of audio and video diffusion models for joint generation. arXiv preprint arXiv:2409.17550 (2024)

[24] 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7822–7832 (2023)

[25] Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)

[26] 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)

[27] Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022), https://arxiv.org/abs/1312.6114

[28] Kuaishou, I.: Kling video generation (2024), https://klingai.com/

[29] Lee, S.H., Oh, G., Byeon, W., Kim, C., Ryoo, W.J., Yoon, S.H., Cho, H., Bae, J., Kim, J., Kim, S.: Sound-guided semantic video generation. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. pp. 34–50. Springer (2022)

[30] 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)

[31] Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)

[32] 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. arXiv preprint arXiv:2301.12503 (2023)

[33] Liu, H., Lan, G.L., Mei, X., Ni, Z., Kumar, A., Nagaraja, V., Wang, W., Plumbley, M.D., Shi, Y., Chandra, V.: Syncflow: Toward temporally aligned joint audio-video generation from text. arXiv preprint arXiv:2412.15220 (2024)

[34] Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2871–2883 (2024)

[35] Liu, H., Luo, K., Wang, W., Chen, Q., Sun, P., Huang, R., Li, X., Ye, J., Xue, W.: Prismaudio: Decomposed chain-of-thoughts and multi-dimensional rewards for video-to-audio generation (2025), https://arxiv.org/abs/2511.18833

[36] Liu, H., Wang, J., Luo, K., Wang, W., Chen, Q., Zhao, Z., Xue, W.: Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing (2025), https://arxiv.org/abs/2506.21448

[37] 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)

[38] Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation (2025), https://arxiv.org/abs/2510.01284

[39] Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)

[40] 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)

[41] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html

[42] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597

[43] 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: CVPR (2023)

[44] Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)

[45] Sun, M., Wang, W., Qiao, Y., Sun, J., Qin, Z., Guo, L., Zhu, X., Liu, J.: Mm-ldm: Multi-modal latent diffusion model for sounding video generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10853–10861 (2024)

[46] Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems **36**, 16083–16099 (2023)

[47] 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)

[48] 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)

[49] Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report (2023), https://arxiv.org/abs/2308.06571

[50] Wang, J., Zeng, X., Qiang, C., Chen, R., Wang, S., Wang, L., Zhou, W., Cai, P., Zhao, J., Li, N., et al.: Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. arXiv preprint arXiv:2506.19774 (2025)

[51] Wang, K., Deng, S., Shi, J., Hatzinakos, D., Tian, Y.: Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. arXiv preprint arXiv:2406.07686 (2024)

[52] Wang, X., Song, R., Li, C., Cheng, X., Li, B., Wu, Y., Wang, Y., Xu, H., Wang, Y.: Animate and sound an image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23369–23378 (June 2025)

[53] Wang, X., Wang, Y., Wu, Y., Song, R., Tan, X., Chen, Z., Xu, H., Sui, G.: Tiva: Time-aligned video-to-audio generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 573–582 (2024)

[54] Weijie Kong, Q.T.: Hunyuanvideo: A systematic framework for large video generative models (2024), https://arxiv.org/abs/2412.03603

[55] Weng, S., Zheng, H., Chang, Z., Li, S., Shi, B., Wang, X.: Audio-sync video generation with multi-stream temporal control. arXiv preprint arXiv:2506.08003 (2025)

[56] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324

[57] 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)

[58] Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)

[59] 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)

[60] Zhang, L., Mo, S., Zhang, Y., Morgado, P.: Audio-synchronized visual animation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)

[61] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)

[62] Zhao, L., Feng, L., Ge, D., Chen, R., Yi, F., Zhang, C., Zhang, X.L., Li, X.: Uniform: A unified multi-task diffusion transformer for audio-video generation. arXiv preprint arXiv:2502.03897 (2025)

[63] Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)