Novelty Assessment Report

Paper: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization

PDF URL: https://openreview.net/pdf?id=y7HV7KT3Bd

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

This paper introduces JavisDiT, a novel Joint Audio-Video Diffusion Transformer designed for synchronized audio-video generation (JAVG). Based on the powerful Diffusion Transformer (DiT) architecture, JavisDiT simultaneously generates high-quality audio and video content from open-ended user prompts in a unified framework. To ensure audio-video synchronization, we introduce a fine-grained spatio-temporal alignment mechanism through a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator. This module extracts both global and fine-grained spatio-temporal priors, guiding the synchronization between the visual and auditory components. Furthermore, we propose a new benchmark, JavisBench, which consists of 10,140 high-quality text-captioned sounding videos and focuses on synchronization evaluation in diverse and complex real-world scenarios. Further, we specifically devise a robust metric for measuring the synchrony between generated audio-video pairs in real-world content. Experimental results demonstrate that JavisDiT significantly outperforms existing methods by ensuring both high-quality generation and precise synchronization, setting a new standard for JAVG tasks.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **synchronized audio-video generation from text prompts**

A total of **50 papers** were analyzed and organized into a taxonomy with **17 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Unified Joint Audio-Video Generation Architectures**
- **Multi-Stage and Cascaded Generation Pipelines**
- **Speech-Driven Talking Head Synthesis**
- **Video-Conditioned Audio Generation**
- **Synchronization and Alignment Mechanisms**
- **Benchmarking and Evaluation Frameworks**
- **Application-Specific Generation Systems**
- **Retrieval and Cross-Modal Understanding**
- **Long-Form and Infinite Generation**
- **Theoretical and Architectural Surveys**

Complete Taxonomy Tree

- synchronized audio-video generation from text prompts Survey Taxonomy
- Unified Joint Audio-Video Generation Architectures
  - Dual-Branch Diffusion Transformer Architectures ★ (8 papers)
  - [0] Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization (Anon et al., 2026) View paper
  - [3] Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions (Zhang Guozhen, 2025) View paper
  - [5] SyncFlow: Toward Temporally Aligned Joint Audio-Video Generation from Text (Liu, 2024) View paper
  - [14] Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation (Wang Kai, 2024) View paper
  - [27] Av-dit: Taming image diffusion transformers for efficient joint audio and video generation (Kai Wang, 2025) View paper
  - [41] JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization (Liu Kai, 2025) View paper
  - [43] 3MDiT: Unified Tri-Modal Diffusion Transformer for Text-Driven Synchronized Audio-Video Generation (Yaoru Li, 2025) View paper
  - [45] ProAV-DiT: A Projected Latent Diffusion Transformer for Efficient Synchronized Audio-Video Generation (Jiahui Sun, 2025) View paper
  - Shared-Backbone Unified Models (4 papers)
  - [13] Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation (Wang Le, 2025) View paper
  - [17] Ovi: Twin backbone cross-modal fusion for audio-video generation (Wang WeiMin, 2025) View paper
  - [32] Sounding video generator: A unified framework for text-guided sounding video generation (Jiawei Liu, 2023) View paper
  - [50] JoVA: Unified Multimodal Learning for Joint Video-Audio Generation (Xiaohu Huang, 2025) View paper
  - Expert Model Fusion and Adaptation (3 papers)
  - [1] Diverse and aligned audio-to-video generation via text-to-video model adaptation (Itai Gat, 2024) View paper
  - [2] UniVerse-1: Unified Audio-Video Generation via Stitching of Experts (Wang Duo-min, 2025) View paper
  - [26] Language-Guided Joint Audio-Visual Editing via One-Shot Adaptation (Susan Liang, 2024) View paper

- Multi-Stage and Cascaded Generation Pipelines
  - Text-to-Audio-to-Video Cascades (3 papers)
  - [10] Text-to-Audio Generation Synchronized with Videos (Mo, 2024) View paper
  - [11] DiffAVA: Personalized Text-to-Audio Generation with Visual Alignment (Mo, 2023) View paper
  - [12] Syncphony: Synchronized Audio-to-Video Generation with Diffusion Transformers (Song Jibin, 2025) View paper
  - Multi-Modal Conditioning with Intermediate Representations (4 papers)
  - [4] Vintage: Joint video and text conditioning for holistic audio generation (Saksham Singh Kushwaha, 2025) View paper
  - [15] AADiff: Audio-Aligned Video Synthesis with Text-to-Image Diffusion (Lee Seungwoo, 2023) View paper
  - [28] TA2V: Text-Audio Guided Video Generation (Minglu Zhao, 2024) View paper
  - [31] Harmony: Harmonizing Audio and Video Generation through Cross-Task Synergy (Teng Hu, 2025) View paper
- Speech-Driven Talking Head Synthesis
  - Text-to-Talking-Head Generation (6 papers)
  - [7] Anyonenet: Synchronized speech and talking head generation for arbitrary persons (Xinsheng Wang, 2022) View paper
  - [8] Text-Driven Synchronized Diffusion Video and Audio Talking Head Generation (Zhen-fei Zhang, 2024) View paper
  - [9] OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking (Wang Zhong-jian, 2025) View paper
  - [24] AV-Flow: Transforming Text to Audio-Visual Human-like Interactions (Chatziagapi, 2025) View paper
  - [39] Ada-TTA: Towards Adaptive High-Quality Text-to-Talking Avatar Synthesis (Ye, 2023) View paper
  - [46] Text-to-Digital Person Video Generator: DigitalAvatarGen (Kesharwani, 2024) View paper
  - Audio-Driven Facial Animation and Lip Synchronization (5 papers)
  - [6] Audio-synchronized visual animation (Lin Zhang, 2024) View paper
  - [44] JAM-Flow: Joint Audio-Motion Synthesis with Flow Matching (Kwon, 2025) View paper
  - [47] End to end lip synchronization with a temporal autoencoder (Shalev, 2020) View paper
  - [48] Shared Latent Representation for Joint Text-to-Audio-Visual Synthesis (Yaman, 2025) View paper
  - [49] VSpeechLM: A Visual Speech Language Model for Visual Text-to-Speech Task (Yuyue Wang, 2025) View paper
- Video-Conditioned Audio Generation
  - Video-to-Audio Synthesis with Temporal Alignment (2 papers)
  - [18] Audio-Sync Video Generation with Multi-Stream Temporal Control (Weng, 2025) View paper
  - [23] TA-V2A: Textually Assisted Video-to-Audio Generation (Wu Xihong, 2025) View paper
  - Multi-Modal Video-to-Audio Synthesis (2 papers)
  - [19] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation (J Zheng, 2025) View paper
  - [37] TAVID: Text-Driven Audio-Visual Interactive Dialogue Generation (Ji-Hoon Kim, 2025) View paper
- Synchronization and Alignment Mechanisms (4 papers)
  - [20] Transface: Unit-based audio-visual speech synthesizer for talking head translation (Xize Cheng, 2024) View paper
  - [22] Synchronized Speech and Video Synthesis (Aaditya Shivprakash Barve, 2023) View paper
  - [25] Dubwise: Video-guided speech duration control in multimodal llm-based text-to-speech for dubbing (Neha Sahipjohn, 2024) View paper
  - [34] M3TTS: Multi-modal text-to-speech of multi-scale style control for dubbing (Yan Liu, 2024) View paper
- Benchmarking and Evaluation Frameworks (1 papers)
  - [30] Tavgbench: Benchmarking text to audible-video generation (Yuxin Mao, 2024) View paper
- Application-Specific Generation Systems
  - Dubbing and Multilingual Video Translation (1 papers)
  - [21] Whisper, Translate, Speak, Sync: Video Translation for Multilingual Video Conferencing Using Generative AI (Amirkia Rafiei Oskooei, 2025) View paper
  - Presentation and Storytelling Video Synthesis (3 papers)
  - [16] PresentAgent: Multimodal Agent for Presentation Video Generation (Jingwei Shi, 2025) View paper
  - [36] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models (S. Sridhar, 2025) View paper
  - [38] Text-to-visual adaptation: Image and audio synthesis for storytelling (R. P. Ram Kumar, 2025) View paper
  - Automated Content Creation and Conversion Tools (2 papers)
  - [35] TEXT2AV â□□ Automated Text to Audio and Video Conversion (Sanjeeva Polepaka, 2023) View paper
  - [40] Text-to-audiovisual speech synthesizer (Udit Kumar Goyal, 2000) View paper
- Retrieval and Cross-Modal Understanding (1 papers)
  - [29] Audio-Enhanced Text-to-Video Retrieval using Text-Conditioned Feature Alignment (Sarah Ibrahimi, 2023) View paper
- Long-Form and Infinite Generation (1 papers)
  - [33] FLAV: Rolling Flow matching for infinite Audio Video generation (A Ergasti, 2025) View paper
- Theoretical and Architectural Surveys (1 papers)
  - [42] Artificial Intelligence in Multimedia Content Generation: A Review of Audio and Video Synthesis Techniques (C Ding, 2025) View paper

Narrative

Core task: synchronized audio-video generation from text prompts. The field has organized itself around several complementary strategies for producing coherent audiovisual content. Unified joint architectures aim to generate both modalities simultaneously within a single model, often leveraging dual-branch diffusion transformers or shared latent representations to maintain tight synchronization. Multi-stage and cascaded pipelines decompose the problem into sequential steps—first generating one modality, then conditioning the other—while speech-driven talking head synthesis focuses on the specialized case of animating human faces from audio. Video-conditioned audio generation reverses the dependency, producing soundtracks that match visual events, and dedicated synchronization mechanisms ensure temporal alignment across modalities. Benchmarking frameworks and application-specific systems address evaluation and real-world deployment, while retrieval and cross-modal understanding methods explore how to leverage existing audiovisual data. Long-form generation and theoretical surveys round out the taxonomy by tackling scalability and providing conceptual overviews.

Within the unified joint architectures, a particularly active line of work centers on dual-branch diffusion transformers, where separate but interacting branches handle audio and video streams. Joint Audio Video Diffusion[0] exemplifies this approach by employing parallel diffusion pathways with cross-modal attention to ensure frame-level synchronization. Nearby efforts such as AV DiT[14] and AV DiT Taming[27] explore similar dual-branch designs, experimenting with different attention mechanisms and training strategies to balance generation quality against computational cost. In contrast, works like Uniavgen[3] and SyncFlow[5] emphasize tighter integration or flow-based formulations, trading architectural simplicity for potentially stronger alignment guarantees. The main open questions revolve around how much cross-modal interaction is necessary during generation versus post-hoc alignment, and whether fully unified models can match the flexibility of cascaded pipelines without sacrificing synchronization fidelity.

Related Works in Same Category

The following **7 sibling papers** share the same taxonomy leaf node with the original paper:

1. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions

Authors: Zhang Guozhen, Zhou Zixiang, Guozhen Zhang, Hu Teng, Zixiang Zhou, et al. (19 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Due to the lack of effective cross-modal modeling, existing open-source audio-video generation methods often exhibit compromised lip synchronization and insufficient semantic consistency. To mitigate these drawbacks, we propose UniAVGen, a unified framework for joint audio and video generation. UniAVGen is anchored in a dual-branch joint synthesis architecture, incorporating two parallel Diffusion Transformers (DiTs) to build a cohesive cross-modal latent space. At its heart lies an Asymmetric C...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video synthesis with cross-modal interaction mechanisms. They overlap in using DiT-based architectures with bidirectional cross-attention for audio-video alignment and synchronization. However, the original paper (JavisDiT) emphasizes hierarchical spatio-temporal prior estimation through its HiST-Sypo module and introduces a new benchmark (JavisBench), while UniAVGen focuses on asymmetric cross-modal interactions with Face-Aware Modulation and Modality-Aware Classifier-Free Guidance for enhanced lip synchronization and semantic consistency.

2. SyncFlow: Toward Temporally Aligned Joint Audio-Video Generation from Text

Authors: Liu, Haohe, Lan, Gael Le, Haohe Liu, et al. (29 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Video and audio are closely correlated modalities that humans naturally perceive together. While recent advancements have enabled the generation of audio or video from text, producing both modalities simultaneously still typically relies on either a cascaded process or multi-modal contrastive encoders. These approaches, however, often lead to suboptimal results due to inherent information losses during inference and conditioning. In this paper, we introduce SyncFlow, a system that is capable of ...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video generation from text prompts. They share overlapping approaches in using separate video and audio branches with cross-modal interaction mechanisms to achieve synchronized generation. The key differences are: JavisDiT introduces a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator for fine-grained spatio-temporal alignment and proposes the JavisBench benchmark, while SyncFlow focuses on a dual-diffusion-transformer (d-DiT) architecture with a modality adaptor and employs a modality-decoupled multi-stage training strategy for computational efficiency.

3. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation

Authors: Wang Kai, Deng Shi-jian, Kai Wang, Shi Jing, Shijian Deng, et al. (11 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recent Diffusion Transformers (DiTs) have shown impressive capabilities in generating high-quality single-modality content, including images, videos, and audio. However, it is still under-explored whether the transformer-based diffuser can efficiently denoise the Gaussian noises towards superb multimodal content creation. To bridge this gap, we introduce AV-DiT, a novel and efficient audio-visual diffusion transformer designed to generate high-quality, realistic videos with both visual and audio...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video synthesis. They overlap in using DiT-based architectures with cross-modal interaction mechanisms to achieve synchronized generation from text prompts. The key difference is that the original paper (JavisDiT) introduces a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator for fine-grained spatio-temporal alignment and proposes a new benchmark (JavisBench), while AV-DiT focuses on parameter-efficient adaptation by freezing a pre-trained image DiT backbone and adding lightweight trainable adapters (LoRA, temporal adapters) to enable joint generation with minimal computational cost.

4. Av-dit: Taming image diffusion transformers for efficient joint audio and video generation

Authors: Kai Wang, Shi-Jian Deng, Jing Shi, Shijian Deng, Dimitrios Hatzinakos, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent Diffusion Transformers (DiTs) have shown impressive capabilities in generating single-modality content, including images, videos, and audio. However, the potential of DiTs to enable superb multimodal content creation remains underexplored. To bridge this gap, we introduce AV-DiT, a novel and efficient audio-visual diffusion transformer designed to generate high-quality, realistic videos with synchronized audio tracks. To minimize model complexity and computational costs, our AV-DiT utiliz...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video generation from text prompts. While both share the approach of using separate but interacting branches with cross-modal attention mechanisms, the original paper (JavisDiT) emphasizes hierarchical spatio-temporal prior synchronization through a dedicated HiST-Sypo Estimator module and introduces a new benchmark (JavisBench) with robust evaluation metrics, whereas AV-DiT focuses on parameter efficiency by utilizing a modality-shared DiT backbone pre-trained on image data with only newly inserted adapters being trainable, minimizing model complexity and computational costs.

5. JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization

Authors: Liu Kai, Li Wei, Chen Lai, Wu, Shengqiong, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

This paper introduces JavisDiT, a novel Joint Audio-Video Diffusion Transformer designed for synchronized audio-video generation (JAVG). Built upon the powerful Diffusion Transformer (DiT) architecture, JavisDiT is able to generate high-quality audio and video content simultaneously from open-ended user prompts. To ensure optimal synchronization, we introduce a fine-grained spatio-temporal alignment mechanism through a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator. This ...

⚠ Similarity Notice

These papers share nearly identical titles, abstracts, technical approaches, and core contributions (JavisDiT architecture, HiST-Sypo Estimator, JavisBench benchmark). The content, methodology, figures, and experimental results are essentially the same, indicating these are likely the same paper or very close variants (e.g., conference submission vs. arXiv version).

6. 3MDiT: Unified Tri-Modal Diffusion Transformer for Text-Driven Synchronized Audio-Video Generation

Authors: Yaoru Li, Heyu Si, Federico Landi, Pilar Oplustil Gallegos, Ioannis Koutsoumpas, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Text-to-video (T2V) diffusion models have recently achieved impressive visual quality, yet most systems still generate silent clips and treat audio as a secondary concern. Existing audio-video generation pipelines typically decompose the task into cascaded stages, which accumulate errors across modalities and are trained under separate objectives. Recent joint audio-video generators alleviate this issue but often rely on dual-tower architectures with ad-hoc cross-modal bridges and static, single...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video synthesis from text prompts. They overlap in using DiT-based architectures with cross-modal interaction mechanisms to achieve synchronized generation. However, JavisDiT focuses on hierarchical spatio-temporal prior estimation (HiST-Sypo) with fine-grained spatial and temporal alignment modules, while 3MDiT emphasizes a unified tri-modal architecture with omni-blocks for feature-level fusion and dynamic text conditioning that evolves jointly with audio-video streams.

7. ProAV-DiT: A Projected Latent Diffusion Transformer for Efficient Synchronized Audio-Video Generation

Authors: Jiahui Sun, Weining Wang, Mingzhen Sun, Yirong Yang, Xinxin Zhu, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Sounding Video Generation (SVG) remains a challenging task due to the inherent structural misalignment between audio and video, as well as the high computational cost of multimodal data processing. In this paper, we introduce ProAV-DiT, a Projected Latent Diffusion Transformer designed for efficient and synchronized audio-video generation. To address structural inconsistencies, we preprocess raw audio into video-like representations, aligning both the temporal and spatial dimensions between audi...

Relationship Analysis

Both papers belong to the Dual-Branch Diffusion Transformer Architectures category, employing parallel diffusion transformer branches for joint audio-video synthesis from text prompts. They overlap in using DiT-based architectures with cross-modal attention mechanisms to achieve audio-video synchronization. However, the original paper (JavisDiT) focuses on hierarchical spatio-temporal prior estimation through a dedicated HiST-Sypo module and introduces a new benchmark (JavisBench), while the candidate paper (ProAV-DiT) emphasizes computational efficiency through projected latent representations using orthogonal decomposition (MDSA) and multi-scale attention mechanisms without explicit prior estimation modules.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["JavisDiT introduces a joint audio-video diffusion transformer featuring a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator that extracts multi-level spatio-temporal priors to guide synchronization. The paper resides in the Dual-Branch Diffusion Transformer Architectures leaf, which contains eight papers—a moderately populated research direction within the broader unified joint generation category. This leaf focuses on parallel diffusion pathways with cross-modal interaction, distinguishing itself from single-backbone shared models and multi-stage cascaded pipelines that handle audio and video sequentially.",

"The taxonomy tree reveals that Dual-Branch Diffusion Transformer Architectures sits alongside Shared-Backbone Unified Models (four papers) and Expert Model Fusion (three papers) within the parent category of Unified Joint Audio-Video Generation Architectures. Neighboring branches include Multi-Stage Cascaded Pipelines and Speech-Driven Talking Head Synthesis, which address related but distinct problems. The scope note clarifies that dual-branch methods employ separate but interacting branches, while shared-backbone approaches use a single architecture with modality-specific adapters. JavisDiT's hierarchical prior estimator differentiates it from sibling works that rely primarily on cross-attention or flow-based formulations for alignment.",

"Among 30 candidates examined across three contributions, no clearly refutable prior work was identified. The JavisDiT architecture contribution examined 10 candidates with none refuting the hierarchical spatio-temporal prior mechanism. The JavisBench dataset contribution also examined 10 candidates, finding no existing benchmark specifically targeting synchronization evaluation in diverse real-world scenarios at comparable scale. The JavisScore metric contribution similarly examined 10 candidates without encountering a prior metric explicitly designed for measuring real-world audio-video synchrony. These statistics suggest that within the limited search scope, the paper's specific combination of hierarchical prior extraction, benchmark design, and evaluation metric appears distinct from examined prior work.",

"Given the moderately populated leaf and the limited 30-candidate search, the work appears to introduce novel components—particularly the HiST-Sypo mechanism and synchronized evaluation infrastructure—within an active research direction. However, the analysis does not cover the full corpus of dual-branch diffusion architectures or exhaustively compare against all synchronization mechanisms in neighboring leaves. The absence of refutable candidates reflects the search scope rather than definitive field-wide novelty, and a broader literature review might reveal closer antecedents or parallel developments in related branches."] } ```

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: JavisDiT: Joint Audio-Video Diffusion Transformer with HiST-Sypo Estimator

Description: The authors propose JavisDiT, a diffusion transformer architecture for joint audio-video generation that incorporates a Hierarchical Spatial-Temporal Synchronized Prior (HiST-Sypo) Estimator. This module extracts global coarse-grained and fine-grained spatio-temporal priors from text prompts to guide precise synchronization between generated audio and video content.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MagicTryOn: Harnessing Diffusion Transformer for Garment-Preserving Video Virtual Try-on

URL: View paper

Brief Assessment

MagicTryOn[70] focuses on video virtual try-on for garment synthesis, not joint audio-video generation. The candidate uses diffusion transformers for a completely different application domain (fashion/garment fitting) without any audio generation component or hierarchical spatio-temporal synchronization mechanisms for audio-video alignment.

2. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation

URL: View paper

Brief Assessment

Audiogen Omni[13] focuses on generating audio, speech, and song from video using multimodal diffusion transformers with lyrics-transcription encoding and phase-aligned positional embeddings. It does not address the hierarchical spatial-temporal synchronized prior estimation mechanism that is central to the original paper's contribution.

3. HunyuanVideo-Foley: Multimodal Diffusion with Representation Alignment for High-Fidelity Foley Audio Generation

URL: View paper

Brief Assessment

HunyuanVideo-Foley[73] focuses on video-to-audio generation with representation alignment strategies, not joint audio-video generation with hierarchical spatio-temporal synchronization mechanisms as in the original paper.

4. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation

URL: View paper

Brief Assessment

AV-DiT[14] focuses on parameter-efficient adaptation of pre-trained image DiT for joint audio-video generation using lightweight adapters, without the hierarchical spatio-temporal prior estimation mechanism that is central to the original paper's contribution.

5. DiVE: Efficient Multi-View Driving Scenes Generation Based on Video Diffusion Transformer

URL: View paper

Brief Assessment

DiVE[72] focuses on multi-view driving scene video generation with spatial consistency across camera views, not joint audio-video generation with temporal synchronization between modalities.

6. 360-degree Human Video Generation with 4D Diffusion Transformer

URL: View paper

Brief Assessment

360 Degree Human Video[71] focuses on 360-degree human video generation from a single image using 4D transformers for spatial-temporal coherence across viewpoints. This is fundamentally different from JavisDiT's joint audio-video generation with hierarchical spatio-temporal synchronization between audio and visual modalities.

7. JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization

URL: View paper

Brief Assessment

JavisDiT[41] focuses on joint audio-video generation with hierarchical spatio-temporal synchronization, which is a different technical approach from the original paper's general RL framework contributions.

8. Av-dit: Taming image diffusion transformers for efficient joint audio and video generation

URL: View paper

Brief Assessment

AV DiT Taming[27] focuses on efficient parameter adaptation using a shared image-based DiT backbone for audio-video generation, while the original paper proposes a hierarchical spatio-temporal prior (HiST-Sypo) estimator for fine-grained synchronization guidance—a distinct architectural contribution not present in the candidate.

9. Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis

URL: View paper

Brief Assessment

Snap Video[69] focuses exclusively on text-to-video generation using spatiotemporal transformers, without any audio generation or audio-video synchronization components. It does not address joint audio-video generation tasks.

10. Tora: Trajectory-oriented Diffusion Transformer for Video Generation

URL: View paper

Brief Assessment

Tora[68] focuses on trajectory-oriented video generation with motion control, not joint audio-video generation with hierarchical spatio-temporal synchronization between audio and video modalities.

Contribution 2: JavisBench: A challenging benchmark dataset for joint audio-video generation

Description: The authors introduce JavisBench, a benchmark consisting of 10,140 high-quality text-captioned sounding videos spanning 5 dimensions and 19 scene categories. The dataset emphasizes complex multi-event scenarios with diverse spatial and temporal compositions to enable comprehensive evaluation of joint audio-video generation systems in real-world contexts.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Mecd: Unlocking multi-event causal discovery in video reasoning

URL: View paper

Brief Assessment

MECD[67] focuses on multi-event causal discovery in video reasoning, not joint audio-video generation. The datasets serve fundamentally different purposes: MECD analyzes causal relationships between events in videos, while JavisBench evaluates synchronized audio-video generation quality.

2. Fine-grained audioâ□□visual event localization

URL: View paper

Brief Assessment

Fine Grained Event Localization[64] focuses on fine-grained audio-visual event localization (recognizing and localizing events in videos), not joint audio-video generation from text. The IT-AVE dataset is designed for event detection tasks, not generative modeling evaluation.

3. SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation

URL: View paper

Brief Assessment

SAVGBench[60] focuses on spatially aligned audio-video generation with stereo/multichannel audio and spatial sound events, while JavisBench emphasizes temporal synchronization and multi-event scenarios without spatial audio considerations. These are distinct research directions within audio-video generation.

4. Temporally Aligned Audio for Video with Autoregression

URL: View paper

Brief Assessment

Temporally Aligned Audio[52] focuses on video-to-audio generation with a benchmark (VisualSound) derived from VGGSound, not joint audio-video generation from text prompts with multi-event scenarios.

5. Toward long form audio-visual video understanding

URL: View paper

Brief Assessment

Long Form Audio Visual[66] focuses on multisensory temporal event localization in long videos (average 210 seconds), not joint audio-video generation from text prompts. The tasks and objectives are fundamentally different.

6. video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models

URL: View paper

Brief Assessment

video SALMONN[65] focuses on audio-visual understanding and speech comprehension using large language models, not on joint audio-video generation. The candidate introduces a speech-audio-visual evaluation benchmark for video-QA tasks, which is fundamentally different from the ORIGINAL's JavisBench designed to evaluate generative models for synchronized audio-video creation.

7. Audio-Sync Video Generation with Multi-Stream Temporal Control

URL: View paper

Brief Assessment

Audio Sync Video[18] focuses on audio-sync video generation using demixed audio tracks (speech, effects, music) with the DeMix dataset, not on joint audio-video generation benchmarks. The candidate's dataset is designed for controllable video generation from audio inputs, while the original contribution emphasizes evaluation of joint generation systems with multi-event scenarios.

8. Perception Test: A Diagnostic Benchmark for Multimodal Video Models

URL: View paper

Brief Assessment

Perception Test[62] focuses on evaluating multimodal video understanding models through perception and reasoning tasks (memory, abstraction, physics, semantics), not on benchmarking joint audio-video generation systems. The datasets serve fundamentally different purposes in distinct research domains.

9. DAVE: Diagnostic benchmark for Audio Visual Evaluation

URL: View paper

Brief Assessment

DAVE[61] focuses on evaluating audio-visual understanding and question-answering capabilities in multimodal LLMs, not on benchmarking joint audio-video generation systems. The tasks and evaluation paradigms are fundamentally different.

10. FoleyBench: A Benchmark For Video-to-Audio Models

URL: View paper

Brief Assessment

FoleyBench[63] focuses specifically on video-to-audio (V2A) generation for foley sound effects, not joint audio-video generation. The original paper's benchmark evaluates simultaneous generation of both modalities from text, which is a fundamentally different task.

Contribution 3: JavisScore: A robust metric for audio-video synchronization evaluation

Description: The authors develop JavisScore, a new evaluation metric based on temporal-aware semantic alignment that measures spatio-temporal synchronization in diverse real-world scenarios. This metric addresses limitations of existing metrics like A V-Align by using a windowed approach with ImageBind encoders to assess audio-visual alignment across video segments.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Audioscenic: Audio-driven video scene editing

URL: View paper

Brief Assessment

AudioScenic[58] focuses on audio-driven video scene editing (background manipulation) rather than joint audio-video generation. While both papers use temporal windowing with semantic alignment for evaluation, AudioScenic's metric measures editing quality in a different task context (scene editing vs. synchronized generation), making direct novelty refutation unclear.

2. TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining

URL: View paper

Brief Assessment

TACOS[56] focuses on temporally-aligned audio captions for language-audio pretraining, not on audio-video synchronization metrics. The candidate addresses audio-text alignment rather than audio-visual synchronization measurement.

3. STA-V2A: Video-to-Audio Generation with Semantic and Temporal Alignment

URL: View paper

Brief Assessment

STA V2A[51] proposes 'audio-audio align' for assessing audio-temporal alignment in video-to-audio generation, which differs from JavisScore's temporal-aware semantic alignment approach using windowed ImageBind encoders for spatio-temporal synchronization evaluation.

4. Smooth-Foley: Creating Continuous Sound for Video-to-Audio Generation Under Semantic Guidance

URL: View paper

Brief Assessment

Smooth Foley[57] focuses on video-to-audio generation with semantic guidance from textual labels, not on developing evaluation metrics for audio-video synchronization. The paper does not propose or discuss synchronization metrics.

5. SpecMaskFoley: Steering Pretrained Spectral Masked Generative Transformer Toward Synchronized Video-to-audio Synthesis via ControlNet

URL: View paper

Brief Assessment

SpecMaskFoley[55] focuses on video-to-audio synthesis using ControlNet and does not propose a new evaluation metric for audio-video synchronization. The paper uses existing metrics like desync for evaluation.

6. Temporally Aligned Audio for Video with Autoregression

URL: View paper

Brief Assessment

Temporally Aligned Audio[52] addresses temporal alignment in video-to-audio generation but does not propose a general metric for evaluating audio-video synchronization across diverse generation tasks.

7. Video-to-Audio Generation with Hidden Alignment

URL: View paper

Brief Assessment

Video to Audio Alignment[53] focuses on video-to-audio generation tasks and evaluates alignment using existing metrics like AV-Align and CAVP, rather than proposing a new temporal-aware semantic alignment metric for evaluation.

8. Towards Video to Piano Music Generation with Chain-of-Perform Support Benchmarks

URL: View paper

Brief Assessment

Video to Piano[59] focuses on piano music generation from video with chain-of-perform guidance, not general audio-video synchronization metrics. The candidate addresses a specialized domain (piano performance) rather than the broad temporal-aware semantic alignment metric proposed in the original paper.

9. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds

URL: View paper

Brief Assessment

FoleyCrafter[54] focuses on video-to-audio generation using semantic adapters and temporal controllers, not on developing evaluation metrics for audio-video synchronization.

10. Text-to-Audio Generation Synchronized with Videos

URL: View paper

Brief Assessment

Text to Audio Video[10] focuses on text-to-audio generation aligned with videos using visual-aligned text embeddings and temporal attention, not on developing evaluation metrics for audio-video synchronization. The candidate's metrics evaluate their own model's visual alignment and temporal consistency, which is a different technical approach from JavisScore's temporal-aware semantic alignment using windowed ImageBind encoders.

Appendix: Text Similarity Detection

Textual similarity detection checked 33 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization

Detected in: Core Task (sibling), Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization View paper
- [1] Diverse and aligned audio-to-video generation via text-to-video model adaptation View paper
- [2] UniVerse-1: Unified Audio-Video Generation via Stitching of Experts View paper
- [3] Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions View paper

- [4] Vintage: Joint video and text conditioning for holistic audio generation View paper
- [5] SyncFlow: Toward Temporally Aligned Joint Audio-Video Generation from Text View paper
- [6] Audio-synchronized visual animation View paper
- [7] Anyonenet: Synchronized speech and talking head generation for arbitrary persons View paper
- [8] Text-Driven Synchronized Diffusion Video and Audio Talking Head Generation View paper
- [9] OmniTalker: One-shot Real-time Text-Driven Talking Audio-Video Generation With Multimodal Style Mimicking View paper
- [10] Text-to-Audio Generation Synchronized with Videos View paper
- [11] DiffAVA: Personalized Text-to-Audio Generation with Visual Alignment View paper
- [12] Syncphony: Synchronized Audio-to-Video Generation with Diffusion Transformers View paper
- [13] Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation View paper
- [14] Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation View paper
- [15] AADiff: Audio-Aligned Video Synthesis with Text-to-Image Diffusion View paper
- [16] PresentAgent: Multimodal Agent for Presentation Video Generation View paper
- [17] Ovi: Twin backbone cross-modal fusion for audio-video generation View paper
- [18] Audio-Sync Video Generation with Multi-Stream Temporal Control View paper
- [19] Aligning What Matters: Masked Latent Adaptation for Text-to-Audio-Video Generation View paper
- [20] Transface: Unit-based audio-visual speech synthesizer for talking head translation View paper
- [21] Whisper, Translate, Speak, Sync: Video Translation for Multilingual Video Conferencing Using Generative AI View paper
- [22] Synchronized Speech and Video Synthesis View paper
- [23] TA-V2A: Textually Assisted Video-to-Audio Generation View paper
- [24] AV-Flow: Transforming Text to Audio-Visual Human-like Interactions View paper
- [25] Dubwise: Video-guided speech duration control in multimodal llm-based text-to-speech for dubbing View paper
- [26] Language-Guided Joint Audio-Visual Editing via One-Shot Adaptation View paper
- [27] Av-dit: Taming image diffusion transformers for efficient joint audio and video generation View paper
- [28] TA2V: Text-Audio Guided Video Generation View paper
- [29] Audio-Enhanced Text-to-Video Retrieval using Text-Conditioned Feature Alignment View paper
- [30] Tavgbench: Benchmarking text to audible-video generation View paper
- [31] Harmony: Harmonizing Audio and Video Generation through Cross-Task Synergy View paper
- [32] Sounding video generator: A unified framework for text-guided sounding video generation View paper
- [33] FLAV: Rolling Flow matching for infinite Audio Video generation View paper
- [34] M3TTS: Multi-modal text-to-speech of multi-scale style control for dubbing View paper
- [35] TEXT2AV â Automated Text to Audio and Video Conversion View paper
- [36] Multimodal Cinematic Video Synthesis Using Text-to-Image and Audio Generation Models View paper
- [37] TAVID: Text-Driven Audio-Visual Interactive Dialogue Generation View paper
- [38] Text-to-visual adaptation: Image and audio synthesis for storytelling View paper
- [39] Ada-TTA: Towards Adaptive High-Quality Text-to-Talking Avatar Synthesis View paper
- [40] Text-to-audiovisual speech synthesizer View paper
- [41] JavisDiT: Joint Audio-Video Diffusion Transformer with Hierarchical Spatio-Temporal Prior Synchronization View paper
- [42] Artificial Intelligence in Multimedia Content Generation: A Review of Audio and Video Synthesis Techniques View paper
- [43] 3MDiT: Unified Tri-Modal Diffusion Transformer for Text-Driven Synchronized Audio-Video Generation View paper
- [44] JAM-Flow: Joint Audio-Motion Synthesis with Flow Matching View paper
- [45] ProAV-DiT: A Projected Latent Diffusion Transformer for Efficient Synchronized Audio-Video Generation View paper
- [46] Text-to-Digital Person Video Generator: DigitalAvatarGen View paper
- [47] End to end lip synchronization with a temporal autoencoder View paper
- [48] Shared Latent Representation for Joint Text-to-Audio-Visual Synthesis View paper
- [49] VSpeechLM: A Visual Speech Language Model for Visual Text-to-Speech Task View paper
- [50] JoVA: Unified Multimodal Learning for Joint Video-Audio Generation View paper
- [51] STA-V2A: Video-to-Audio Generation with Semantic and Temporal Alignment View paper
- [52] Temporally Aligned Audio for Video with Autoregression View paper
- [53] Video-to-Audio Generation with Hidden Alignment View paper
- [54] Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds View paper
- [55] SpecMaskFoley: Steering Pretrained Spectral Masked Generative Transformer Toward Synchronized Video-to-audio Synthesis via ControlNet View paper
- [56] TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining View paper
- [57] Smooth-Foley: Creating Continuous Sound for Video-to-Audio Generation Under Semantic Guidance View paper
- [58] Audioscenic: Audio-driven video scene editing View paper
- [59] Towards Video to Piano Music Generation with Chain-of-Perform Support Benchmarks View paper
- [60] SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation View paper
- [61] DAVE: Diagnostic benchmark for Audio Visual Evaluation View paper
- [62] Perception Test: A Diagnostic Benchmark for Multimodal Video Models View paper
- [63] FoleyBench: A Benchmark For Video-to-Audio Models View paper
- [64] Fine-grained audioâvisual event localization View paper
- [65] video-SALMONN: Speech-Enhanced Audio-Visual Large Language Models View paper
- [66] Toward long form audio-visual video understanding View paper
- [67] Mecd: Unlocking multi-event causal discovery in video reasoning View paper
- [68] Tora: Trajectory-oriented Diffusion Transformer for Video Generation View paper
- [69] Snap Video: Scaled Spatiotemporal Transformers for Text-to-Video Synthesis View paper
- [70] MagicTryOn: Harnessing Diffusion Transformer for Garment-Preserving Video Virtual Try-on View paper
- [71] 360-degree Human Video Generation with 4D Diffusion Transformer View paper
- [72] DiVE: Efficient Multi-View Driving Scenes Generation Based on Video Diffusion Transformer View paper
- [73] HunyuanVideo-Foley: Multimodal Diffusion with Representation Alignment for High-Fidelity Foley Audio Generation View paper