

MAVIN: Multi-Shot Audio-Visual Generation with Customized Narrative Control

Kaiqi Liu^{1,2*}, Yunyao Mao^{2*}, Ziqi Cai¹, Zheng Geng³, Jing Wang⁴,
Qiulin Wang², Xintao Wang², Pengfei Wan², Kun Gai²,
Shuchen Weng^{1†}, and Boxin Shi^{1†}

¹ Peking University, Beijing, China

² Kling Team, Kuaishou Technology, Beijing, China

³ Institute of Automation, Chinese Academy of Sciences, Beijing, China

⁴ Sun Yat-sen University, Shenzhen, China

liukq04@gmail.com, {shuchenweng, shiboxin}@pku.edu.cn

Abstract. While recent generative models produce high-fidelity videos, they struggle with the complex narrative control required for coherent multi-shot audio-visual generation. Existing methods suffer from temporal misalignment, limited controllability, and incomplete scripting. In this paper, we propose MAVIN, the first framework for multi-shot audio-visual generation with customized narrative control. To resolve temporal misalignment, we propose boundary-aware attention, which leverages hierarchical captions and boundary-aware token routing to render audio-visual elements within their respective temporal boundaries. To improve the controllability for multi-subject scenarios, we propose ID-aware propagation, utilizing identity embeddings and an identity-aware mask to bind specific identities to consistent visual appearances and vocal timbres. To provide comprehensive audio-visual narratives, we present a multi-agent scripting pipeline to transform free-form user inputs into hierarchical captions. Furthermore, we construct MAVINSet, a multi-shot audio-visual dataset for robust training and evaluation. Extensive experiments demonstrate that MAVIN achieves state-of-the-art performance, opening up a new avenue for integrating generative models into professional filmmaking workflows.

Keywords: Audio-Visual Generation · Video Diffusion Models

1 Introduction

Recent video generation models [13, 22, 55, 62, 72] have achieved remarkable progress in generating high-fidelity videos. With data engineering and architectural adaptation, they are widely explored in specialized scenarios, enabling

¹ Kaiqi Liu is with School of Software and Microelectronics, Peking University. Ziqi Cai, Shuchen Weng, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

* Equal contribution. [†] Corresponding authors.

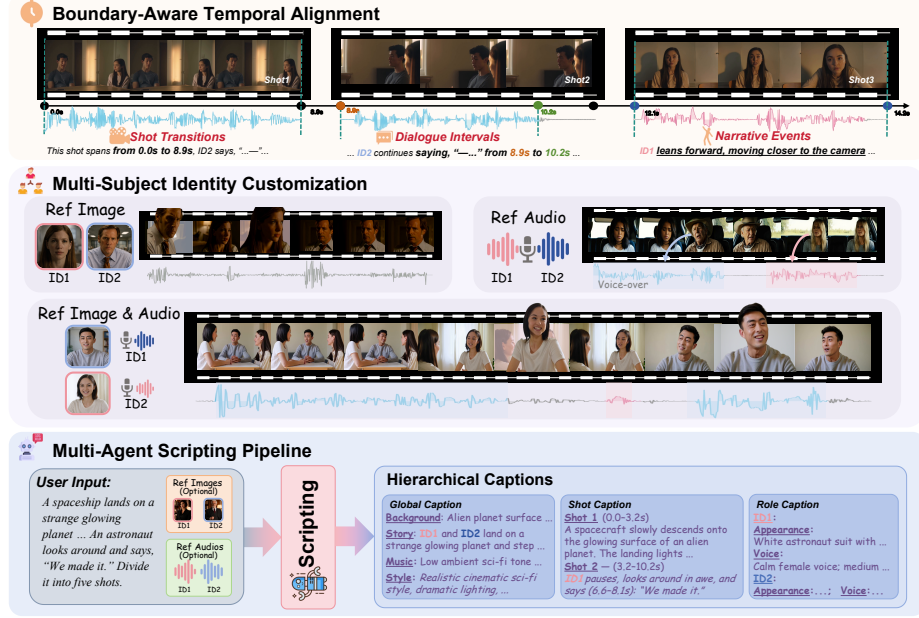


Fig. 1: Illustration of our MAVIN framework. **First row:** Leveraging boundary-aware attention, MAVIN enables precise temporal alignment for shot transitions, dialogue intervals, and narrative events. **Second row:** Through ID-aware propagation, MAVIN allows users to customize multiple subjects via reference images and audio, maintaining identity consistency across complex cinematic narratives. **Third row:** With the multi-agent scripting pipeline, MAVIN translates user inputs into hierarchical captions, providing decoupled narrative semantics for multi-shot audio-visual generation.

precise camera movement [60, 76], long-duration generation [33, 50], and physical dynamics simulation [41, 71]. Towards an immersive experience, researchers increasingly focus on integrating synchronized original audio [18, 39, 51].

However, these clip-level models rely solely on a single text prompt for generation, rendering them impractical for real-world creative workflows (e.g., professional multi-shot filmmaking). In practice, creators depend on storyboards and structured scripts to precisely dictate the video content. To further empower filmmakers, recent efforts have focused on developing agents for story planning [66] and synthesizing keyframes for shot transitions [20, 73], aiming to align with traditional human filmmaking workflows.

Despite the stronger controllability offered by these multi-shot video generation models, previous approaches primarily focus on the video-only modality, struggling with the challenges of audio-visual joint generation across complex narrative timelines for immersive filmmaking: (i) *Temporal misalignment*: Without explicit temporal boundaries, models struggle to synchronize specific visual elements (e.g., a character) with their audio events (e.g., speech) at the correct timestamps. (ii) *Limited controllability*: While current models effectively specify

the visual appearance of subjects, their audio specifications (*e.g.*, vocal timbre) remain largely unexplored, especially in multi-subject scenarios. (iii) *Incomplete scripting*: Prior scripting agents often omit audio-related narratives (*e.g.*, human speech and sound effects), lacking time-stamped audio cues to guide the joint generation process.

To address these challenges, we propose **MAVIN**, the first framework for **M**ulti-**S**hot **A**udio-**V**isual generation with CustomIzed **N**arrative control. We build our framework upon OVI [39] to leverage its priors for joint audio-visual representation and feature alignment. To avoid temporal misalignment, we propose boundary-aware attention, which partitions the latent space and routes corresponding tokens to strictly render audio and visual elements within their respective temporal boundaries (Fig. 1, first row). We further design an ID-aware propagation mechanism to accurately customize user-intended subjects, where ID tokens specify the target characters, and the audio-visual token context is leveraged to control the planned audio-visual attributes (Fig. 1, second row). Finally, a lightweight multi-agent scripting pipeline is introduced to translate free-form prompts into comprehensive audio-visual narratives, providing temporal and structural cues for the joint generation process (Fig. 1, third row).

To facilitate model training and evaluation, we introduce **MAVINSet**, a multi-shot audio-visual dataset with hierarchical captions, constructed from diverse datasets. The training set comprises 800K samples, where each sample ranges from 3–15 seconds in duration and contains 1–6 shots, rendered at 24 fps and 480p resolution. We further provide comprehensive hierarchical annotations at the global, shot, and role levels. These annotations explicitly specify the overall storyline, plot progression, dialogue intervals, action events, character appearance, and vocal timbre. Finally, we establish an evaluation set of 1K manually verified samples to benchmark relevant methods.

Our contributions can be summarized as follows:

- We propose the first framework for multi-shot audio-visual generation with customized narrative control, and introduce a multi-agent scripting pipeline to provide comprehensive audio-visual narratives as guidance.
- We design boundary-aware attention to render audio-visual elements within their respective temporal boundaries, and present an ID-aware propagation mechanism to specify target characters with planned audio-visual attributes.
- We construct a multi-shot audio-visual dataset with customized narrative annotations to facilitate model training, and establish a manually verified benchmark to rigorously evaluate the performance of relevant methods.

2 Related Work

2.1 Joint Audio-Visual Generation

Early audio-visual generation predominantly followed cascaded pipelines (*e.g.*, video-to-audio [40, 80] or audio-to-video [4]), suffering from accumulated cross-modal errors and weak temporal synchronization. Recent research has increasingly shifted toward joint audio-visual generation. Beyond early coupled-diffusion

formulations such as MM-Diffusion [44], methods like Seeing and Hearing [69] bridge pretrained audio and video generators through latent alignment. Concurrently, MM-LDM [49] and AV-DiT [58] move toward more unified latent and transformer-based architectures. More recent DiT-based systems, including UniForm [75], JarvisDiT [36], JarvisDiT++ [37] and OVI [39], further strengthen cross-modal fusion and fine-grained synchronization through shared or twin-backbone designs. Adjacent cross-modal works such as AV-Link [19], MMAudio [8], MTV [61], and VinTAGe [29] further demonstrate the importance of temporally aligned multimodal conditioning for synchronized audio-visual generation. However, these methods largely assume a single contiguous event guided by a global prompt condition. Consequently, they still offer limited support for narrative-level planning and multi-shot semantic transitions.

2.2 Multi-Shot and Narrative Video Generation

To move beyond single-event clips, recent video generation research has begun to explicitly model multi-scene and long-horizon narratives. Planning-oriented methods such as VideoDirectorGPT [34], VideoStudio [38], and STAGE [73] leverage LLMs, spatial layouts, storyboards, or reference images to decompose a global prompt into multiple scenes while preserving cross-scene consistency. Hierarchical frameworks like MovieDreamer [74] further combine long-range autoregressive planning with diffusion rendering for coherent storytelling. At the generation level, StreamingT2V [21] propagates context across extended sequences, while multi-shot models such as ShotAdapter [27] and EchoShot [57] introduce explicit shot-aware conditioning to improve cross-shot controllability and identity preservation. Notably, Mind the Time [68] utilizes temporally localized captions with explicit start and end timestamps, enabling fine-grained control over the duration and ordering of multiple events. Nevertheless, these approaches remain fundamentally video-only, neglecting audio specifications (*e.g.*, vocal timbre) for joint audio-visual generation.

2.3 Controllable Video Generation

Video controllability has rapidly evolved from basic structural conditioning to richer motion, camera, and identity control. Earlier ControlVideo [76], MotionCtrl [60], and Generative Rendering [2] focus on injecting structural, trajectory, or 4D guidance. Subsequent research extends this paradigm toward stronger motion and subject control. Frameworks including MotionBooth [6], Motion Prompting [15], AnimateAnything [30], and MagicMotion [32] support fine-grained object and camera motion specification under increasingly flexible control formats. In parallel, identity-oriented generation has advanced through multi-subject personalization methods such as Video Alchemist [7], improving visual subject fidelity in multi-entity scenes. Adjacent audio-aware editing and control methods further explore language-based colorization with audio alignment [3], audio-synchronized instance editing [77], and instruction-guided joint

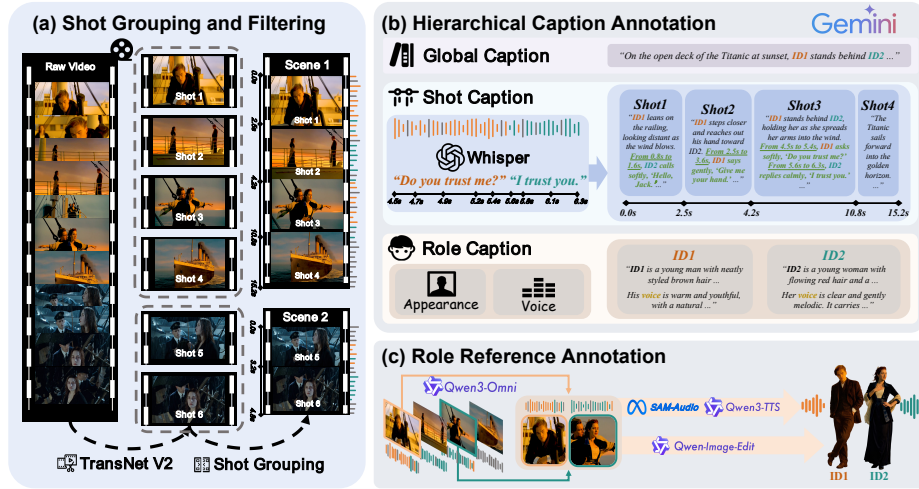


Fig. 2: Overview of our automatic data curation pipeline. **(a) Shot grouping and filtering.** We collect videos from diverse open-source datasets and online platforms. Raw videos are segmented into shots, grouped into multi-shot cinematic clips, and filtered for audio-visual aesthetics and clarity. **(b) Hierarchical caption annotation.** We generate multi-level captions to capture global storylines, shot-level events, and role-level descriptions, alongside dialogue boundaries with decisecond-level timestamps. **(c) Role reference annotation.** To establish consistent identity anchors, we extract audio-visual cues for each role and synthesize perturbed reference images and audio, decoupling identity from exact replication.

audio-video editing [78]. Since existing models primarily customize the visual appearance, we propose MAVIN for audio-visual customized narrative, overcoming temporal misalignment, limited controllability, and incomplete scripting.

3 Dataset

Existing audio-visual datasets [5, 14] mainly focus on single-shot events without structural annotations, leaving a data gap for complex storytelling. To train and evaluate MAVIN and establish a new benchmark for multi-shot generation, we propose MAVINSet, a large-scale dataset annotated with explicit temporal boundaries, dialogue intervals, and multiple audio-visual role references.

Data sources. To ensure both realism and diversity, MAVINSET is built upon diverse in-the-wild resources, including Condensed Movies [1], Short-Films-20K [16], MovieBench [65], VGGSound [5], and OpenHumanVid [31] datasets, and publicly available online videos on YouTube.

Shot unit grouping. As shown in Fig. 2 (a), we process raw videos by enforcing a minimum 480p resolution, removing black borders, and standardizing to 24 fps. We use TransNet V2 [48] for shot boundary detection and apply scene con-

sistency representation learning [64] to group shots belonging to the same scene segment. Each grouped unit includes 1–6 distinct shots spanning 3–15 seconds.

Audio-visual quality filtering. We evaluate grouping candidates using the optical flow from RAFT [52] to discard near-static segments. We then enforce perceptual quality thresholds across modalities, assessing video aesthetics with an aesthetic predictor [45] and audio quality with Audiobox-Aesthetics [53]. To ensure clear identity representation, we use Qwen3-Omni [70] and SAM2 [43] to discard clips with more than three individuals. We also apply Scribe [11] for speaker diarization, excluding clips with interfering multi-speaker audio.

Hierarchical caption annotation. As shown in Fig. 2 (b), we employ Gemini-2.5 [10] to generate hierarchical captions, where the global level captures the overall storyline and plot progression, the role level details character appearance and vocal timbre, and the shot level describes action events and dialogue intervals. Notably, we additionally utilize Whisper [42] to extract timestamps for precise dialogue boundaries. If a dialogue extends beyond a single shot, we append a “—” to indicate continuation.

Role reference annotation. As shown in Fig. 2 (c), for each multi-shot clip group, we extract visual and audio cues using Qwen3-Omni [70] and SAM-Audio [47]. We further synthesize perturbed visual references using Qwen-Image-Edit [63] for altered poses or actions, and audio references using Qwen3-TTS [23] for varying spoken content. This decouples identity from exact replication, enabling the model to generate dynamic actions and speech strictly guided by the narrative script.

Manual benchmark verification. We create a 1K-sample high-fidelity benchmark from MAVINSet for evaluation. To prevent data leakage, we strictly enforce a raw-video-level disjoint split between the training set and evaluation benchmark. We conduct manual verification to discard samples exhibiting identity drift, temporal misalignment, or structural inconsistencies that are beyond the reach of automatic metrics. We further recruit two independent annotators to assess whether each sample should be preserved. This process yielded a high inter-annotator agreement (Cohen’s $\kappa = 0.77$), strongly demonstrating the benchmark’s reliability.

4 Method

In this section, we detail our proposed **MAVIN** framework. We first provide an overview of the model architecture and optimization process (Sec. 4.1). Next, we introduce the boundary-aware attention mechanism to address temporal misalignment across complex narratives (Sec. 4.2). To achieve precise identity and vocal timbre customization, we further detail the ID-aware propagation strategy (Sec. 4.3). Finally, we describe the multi-agent scripting pipeline to transform free-form user prompts into the hierarchical captions (Sec. 4.4). The overall pipeline is shown in Fig. 3.

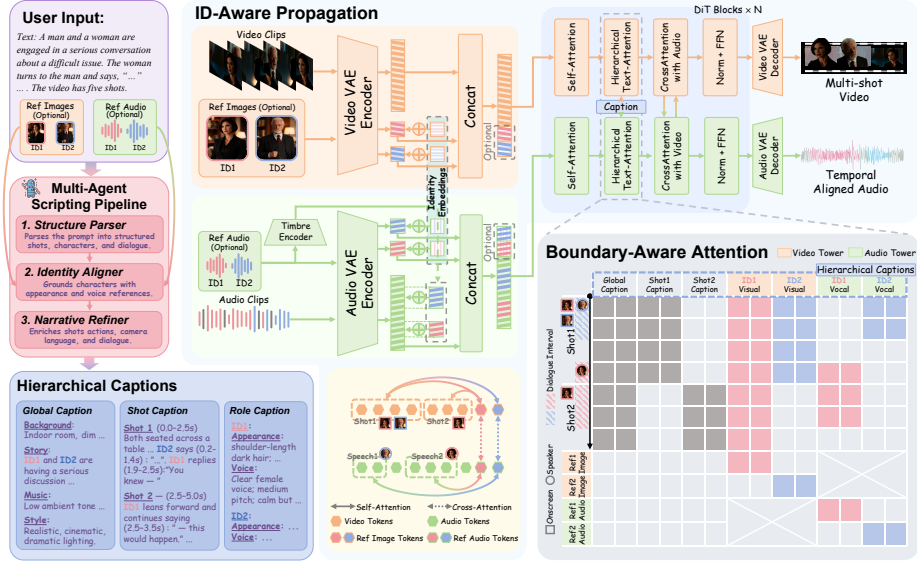


Fig. 3: Given free-form user inputs (**purple block**), we first employ a multi-agent scripting pipeline (**red block**) to transform them into hierarchical captions (**blue block**), detailing complex narrative timelines with precise shot boundaries and dialogue intervals. Within the ID-aware propagation, input video clips and optional image references are compressed by a Video VAE into latent tokens (**orange block**). The audio tower follows a similar process, but its optional audio references are additionally processed by a pretrained timbre encoder to extract clean timbre embeddings (**green block**). After concatenating with learnable identity embeddings, these tokens are fed into a dual-tower diffusion Transformer (**cyan block**). We adapt the self-attention and text-attention layers to be compatible with customized identities (**yellow block**). Furthermore, we integrate boundary-aware attention into these text-attention layers to enforce strict temporal alignment across complex multi-shot narratives (**gray block**). Finally, the VAE decoders produce the synchronized multi-shot video and audio results.

4.1 Overview

In this section, we present an overview of our MAVIN framework.

Audio and video compression. MAVIN employs pretrained Variational Autoencoders (VAEs) to compress audio and video modalities into their latent spaces. Given a video $\mathbf{x}^v$ and its accompanying audio $\mathbf{x}^a$, the respective VAE encoders compress them into compact latent codes $\mathbf{z}_0^v = E_v(\mathbf{x}^v)$ and $\mathbf{z}_0^a = E_a(\mathbf{x}^a)$. **Model architecture.** Following previous works [18, 39, 51], we design our prediction network as a dual-tower diffusion Transformer framework. Given compressed latent codes $\mathbf{z}_t^v$ and $\mathbf{z}_t^a$ at timestep t , hierarchical captions $\mathcal{C}^T$, and optional role references $\mathcal{C}^R$, the framework is equipped with stacked Transformer blocks. These include self-attention layers to extract audio-visual features, hierarchical text-attention layers to decouple multi-level narrative semantics, and

cross-attention layers for fine-grained audio-visual synchronization. The framework conducts joint prediction of the audio-visual velocity fields.

Optimization process. The framework is optimized via flow matching [35], where independent Gaussian priors $\epsilon^v, \epsilon^a \sim \mathcal{N}(0, I)$ are transformed into the target audio-visual latent codes. The linear interpolation paths are defined as $\mathbf{z}_t^v = (1 - t)\mathbf{z}_0^v + t\epsilon^v$ and $\mathbf{z}_t^a = (1 - t)\mathbf{z}_0^a + t\epsilon^a$ for $t \sim \mathcal{U}(0, 1)$. Specifically, the model is optimized to regress the vector fields using the following objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{z}_0^v, \mathbf{z}_0^a, \epsilon^v, \epsilon^a} \left[\|v_\theta^v(\mathbf{z}_t^v, t, \mathcal{C}^T, \mathcal{C}^R) - (\epsilon^v - \mathbf{z}_0^v)\|_2^2 + \|v_\theta^a(\mathbf{z}_t^a, t, \mathcal{C}^T, \mathcal{C}^R) - (\epsilon^a - \mathbf{z}_0^a)\|_2^2 \right], \quad (1)$$

where $v_\theta^v(\dots)$ and $v_\theta^a(\dots)$ are the velocity fields predicted by the video and audio towers, respectively. During inference, the network synthesizes the final synchronized multi-shot audio-visual sequence by solving the probability flow Ordinary Differential Equation (ODE) defined by the predicted velocity fields.

4.2 Boundary-Aware Attention

Clip-level video generation models typically adopt a single text prompt to guide the generation process, where all text tokens are visible to all latent positions. However, for multi-shot generation, this implicit cross-attention inevitably leads to semantic leakage from one shot into another. To address this, we adopt hierarchical captions to decouple global, shot, and role-level narrative semantics, while routing tokens to render elements within their respective temporal boundaries, as presented in Fig. 3 bottom right.

Hierarchical caption formulation. For a multi-shot narrative, we formulate the input caption into a hierarchical format with three levels: (i) *Global caption* ($\mathcal{C}^{\text{global}} = \{\mathbf{g}\}$) presents the overall storyline and plot progress across the entire generated sequence $[0, T]$, where $\mathbf{g}$ denotes the global narrative context; (ii) *Shot caption* ($\mathcal{C}_i^{\text{shot}} = \{(\mathbf{s}_i, \mathcal{T}_i^s)\}$) guides dialogue intervals and action events, strictly anchored to the defined timestamps of the i -th cinematic shot, where $\mathbf{s}_i$ is the description of shot i within the shot interval $\mathcal{T}_i^s$; (iii) *Role caption* ($\mathcal{C}_j^{\text{role}} = \{(\mathbf{d}_j^v, \mathbf{d}_j^a, \mathcal{T}_j^r)\}$) provides fine-grained identity anchoring. The visual appearance descriptor $\mathbf{d}_j^v$ is activated only when role j is present in the shot, while the audio specification descriptor $\mathbf{d}_j^a$ is presented within its dialogue intervals $\mathcal{T}_j^r$. As a result, the hierarchical caption $\mathcal{C}$ is formulated as:

$$\mathcal{C} = \left\{ \{\mathbf{g}\}, \{(\mathbf{s}_i, \mathcal{T}_i^s)\}_{i=1}^S, \{(\mathbf{d}_j^v, \mathbf{d}_j^a, \mathcal{T}_j^r)\}_{j \in \mathcal{R}} \right\}, \quad (2)$$

where S denotes the number of shots and $\mathcal{R}$ specifies the set of characters.

Latent boundary partition. Since videos and audio are presented with distinct physical natures, we propose an asymmetric boundary partition strategy for the latent tokens to construct overlapping segments. Since video frames are continuous, we partition its latent codes $\mathbf{Z}_i^v$ from the video latent space $\mathbf{Z}^v$

according to shot transitions as $\mathbf{Z}_i^v = \{\mathbf{z}_k^v \mid \tau(k) \in \mathcal{T}_i^s\}$. In contrast, since character speech is typically sparse and event-driven, we partition its latent codes $\mathbf{Z}_j^a$ from the audio latent space $\mathbf{Z}^a$ based on the dialogue interval $\mathcal{T}_j^r$ of role j as $\mathbf{Z}_j^a = \{\mathbf{z}_k^a \mid \tau(k) \in \mathcal{T}_j^r\}$.

Boundary-aware token routing. To strictly generate multi-shot videos following this structured input, we replace the standard cross-attention layer with the hierarchical text-attention layer. Denoting a binary routing matrix $\mathbf{M} \in \{0, 1\}^{N \times L}$, where N is the latent sequence length and L is the text token length, the attention calculation is reformulated as:

$$\mathbf{Z} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}_C^\top}{\sqrt{d}} + \log \mathbf{M} \right) \mathbf{V}_C, \quad (3)$$

where $\mathbf{Q}$ is derived from the latent codes, $\mathbf{K}_C$ and $\mathbf{V}_C$ are projected from the hierarchical captions $\mathcal{C}$, d is the head dimension, and the routing matrix $\mathbf{M}$ computes visibility dynamically based on the partitioned temporal boundaries. For a video latent token k at time $\tau(k)$ and a textual token c from the hierarchical caption $\mathcal{C}$, the visibility is defined as:

$$\mathbf{M}^v(k, c) = \begin{cases} 1 & \text{if } c \in \mathcal{C}^{\text{global}}, \\ 1 & \text{if } c \in \mathcal{C}_i^{\text{shot}} \wedge \tau(k) \in \mathcal{T}_i^s, \\ 1 & \text{if } c \in \mathcal{C}_j^{\text{role}} \wedge \tau(k) \in \mathcal{T}_i^s \wedge j \in \mathcal{R}_i, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

where $\mathcal{R}_i$ means the set of characters presented in i -th shot. Similarly, for audio latent codes, the role-level audio specification descriptor $\mathbf{d}_j^a$ is routed to $\mathbf{M}^a(k, c) = 1$ only if $\tau(k) \in \mathcal{T}_j^r$. Finally, we denote the boundary-aware attention mechanism as the collection of latent boundary partition and boundary-aware token routing, enabling rendering audio and visual elements within their respective temporal boundaries.

4.3 ID-Aware Propagation

To improve the controllability of audio-visual generation and enable customized generation for multi-subject scenarios (*i.e.*, binding roles to specific visual appearances and vocal timbres), we introduce identity embeddings for appearance and timbre modulation, and identity-aware injection for customized token conditioning during generation, as presented in Fig. 3 middle.

Identity embeddings. We design identity embeddings for the cases where the user specifies customized roles, each accompanied by optional reference images and audio clips. In this case, for each customized role j , we feed its reference inputs into its respective VAE encoders to obtain tokens representing the visual appearance $\mathbf{Z}_j^{v, \text{ref}}$ and audio specifications $\mathbf{Z}_j^{a, \text{ref}}$. Since reference audio inherently entangles timbre with semantics, emotion, and prosody, we additionally employ a pretrained timbre encoder [26] to extract a clean timbre embedding $\mathbf{e}_j$. Before fusion, we project $\mathbf{e}_j$ to the audio-token hidden dimension and broadcast it to

match the shape of $\mathbf{Z}_j^{\text{a,ref}}$. To explicitly prevent identity collapse among similar roles, we modulate these extracted features with learnable role ID embeddings $\mathbf{p}_j^{\text{v}}, \mathbf{p}_j^{\text{a}} \in \mathbb{R}^d$ as:

$$\hat{\mathbf{Z}}_j^{\text{v,ref}} = \mathbf{Z}_j^{\text{v,ref}} + \mathbf{p}_j^{\text{v}}, \quad \hat{\mathbf{Z}}_j^{\text{a,ref}} = \mathbf{Z}_j^{\text{a,ref}} + \mathbf{p}_j^{\text{a}} + \mathbf{e}_j. \quad (5)$$

These augmented identity anchors serve as customized tokens and are concatenated with their respective target latent codes ($\mathbf{Z}_j^{\text{a}}$ and $\mathbf{Z}_i^{\text{v}}$), anchoring identity priors to the target characters before the diffusion Transformer. If no customized role is specified, this concatenation is omitted.

Identity-aware mask. To prevent the reference images and audio from interfering with the original audio-visual generation process, we apply an additional visibility mask within each attention layer during customization. (i) *Video self-attention layer*: Video tokens $\mathbf{Z}_i^{\text{v}}$ for the i -th shot can only interact with themselves, or with the visual anchors $\hat{\mathbf{Z}}_j^{\text{v,ref}}$ of the roles j present in the same shot. (ii) *Audio self-attention layer*: Audio tokens $\mathbf{Z}_j^{\text{a}}$ can only attend to themselves, or the vocal anchors $\hat{\mathbf{Z}}_j^{\text{a,ref}}$ speaking within their timestamp intervals $\mathcal{T}_j^{\text{r}}$. (iii) *Text-attention layer*: Role-level text prompts strictly attend to their corresponding visual and vocal anchors (i.e., $\mathcal{C}_j^{\text{role}} \leftrightarrow \{\hat{\mathbf{Z}}_j^{\text{v,ref}}, \hat{\mathbf{Z}}_j^{\text{a,ref}}\}$) to propagate accurate role identity. (iv) *Cross-attention layer*: Only tokens belonging to the visual and vocal anchors of the same role are permitted to interact ($\hat{\mathbf{Z}}_j^{\text{a,ref}} \leftrightarrow \hat{\mathbf{Z}}_j^{\text{v,ref}}$). With this carefully designed masking strategy, our model is elegantly adapted for audio-visual role customization.

4.4 Multi-Agent Scripting Pipeline

In our multi-shot audio-visual generation model, the narrative script with hierarchical captions directly decides the final generation quality. Since user-provided prompts $\mathcal{P}$ are typically free-form, lacking fine-grained details for scenario descriptions or role traits, we introduce a lightweight multi-agent scripting pipeline with an off-the-shelf LLM [70] to transform $\mathcal{P}$ into the hierarchical captions $\mathcal{C}$, as presented in Fig. 3 left.

Specifically, we implement this caption augmentation through three specialized agents. First, the *structure parser* takes the user-provided prompt $\mathcal{P}$ and utilizes predefined templates to extract global scene attributes, decomposing each shot boundary $\{\mathcal{T}_i^{\text{s}}\}_{i=1}^S$, identifying the discrete set of characters $\mathcal{R}$, and extracting their dialogue intervals $\{\mathcal{T}_j^{\text{r}}\}_{j \in \mathcal{R}}$. Next, the *identity aligner* grounds these roles by integrating optional user-provided visual and audio references. Finally, the *narrative refiner* enriches the shot-level action events, camera language, and dialogues, producing the final captions $\mathcal{C} = \{\mathcal{C}^{\text{global}}, \{\mathcal{C}_i^{\text{shot}}\}_{i=1}^S, \{\mathcal{C}_j^{\text{role}}\}_{j \in \mathcal{R}}\}$, which are tailored for the formulation of our generative framework.

Table 1: Quantitative comparison of our framework with state-of-the-art audio-visual generation methods. Throughout the paper, $\uparrow$ ($\downarrow$) indicates higher (lower) is better. The best results are highlighted with **bold**, and ‘-’ indicates the metric is inapplicable.

Method	Quality		Semantics			AV-Alignment			Multi-shot Consistency				
	FVD $\downarrow$	FAD $\downarrow$	TVS $\uparrow$	TAS $\uparrow$	WER $\downarrow$	Sync $\uparrow$	AV-IB $\uparrow$	TAMS $\uparrow$	SC $\uparrow$	BC $\uparrow$	CISC $\uparrow$	BISC $\uparrow$	STA $\uparrow$
<i>- Cascaded Generation Models: Multi-shot T2V + V2A</i>													
VideoGen-of-Thought [79]	418.7	13.2	0.1537	0.1939	-	2.278	0.184	-	0.9593	0.9578	0.6195	0.7802	-
MovieAgent [66]	395.3	15.9	0.1271	0.1995	-	1.913	0.179	-	0.9656	0.9298	0.4872	0.6744	-
EchoShot [57]	276.3	12.3	0.2072	0.2191	-	2.095	0.207	-	0.9532	0.9572	0.6104	0.7599	0.6002
CineTrans [67]	254.2	11.9	0.2013	0.2098	-	1.892	0.201	-	0.9425	0.9543	0.6073	0.7603	0.5820
IC-LoRA [24] + Wan [55]	268.5	12.7	0.2193	0.2012	-	2.103	0.192	-	0.9512	0.9626	0.5286	0.7519	-
StoryDiff. [81] + Wan [55]	304.6	13.3	0.2124	0.1978	-	2.013	0.191	-	0.9487	0.9654	0.5613	0.7993	-
<i>- Joint Generation Models: T2AV</i>													
JavisDiT [36]	512.6	21.3	0.163	0.1937	0.305	3.734	0.239	0.1012	0.9512	0.9567	-	-	-
UniVerse-1 [56]	356.9	11.2	0.182	0.1720	0.199	3.851	0.193	0.2892	0.9499	0.9601	-	-	-
OVI [39]	318.4	8.3	0.1983	0.2203	0.093	4.231	0.227	0.4432	0.9565	0.9578	0.6121	0.7789	0.4992
LTX-2 [18]	289.7	8.6	0.2001	0.2287	0.097	4.365	0.203	0.5343	0.9593	0.9553	0.6015	0.7812	0.5733
Ours (MAVIN)	231.6	6.8	0.2471	0.2392	0.048	6.032	0.263	0.8104	0.9695	0.9657	0.6319	0.8013	0.9897

5 Experiments

5.1 Implementation Details

Basic setup. We adopt OVI [39] as the backbone for our joint audio-visual generation, supporting variable clip lengths of 3 to 15 seconds at 24 fps with 480p resolution. Training is performed within a joint latent space encoded by 3D video and audio VAEs. To handle variable sequence lengths, we employ dynamic resolution bucket-cropping and precise temporal alignment strategies.

Multi-stage progressive training. To prevent optimization conflicts caused by diverse conditioning signals (*i.e.*, multi-level narrative semantics, customized visual appearances and vocal timbres), we adopt a progressive alignment training schedule. We sequentially optimize for text-driven generation (Stage 1), introduce single-modality identity anchors for visual and audio references (Stage 2), and finally perform joint training integrating all conditions (Stage 3).

5.2 Evaluation Metrics

We comprehensively evaluate MAVIN from the following perspectives:

Perceptual quality. We measure the fidelity of generated video and audio using Fréchet Video Distance (FVD) [54] and Fréchet Audio Distance (FAD) [28].

Semantic consistency. We employ ViCLIP [59] and CLAP [12] to measure Text-Video Similarity (TVS) and Text-Audio Similarity (TAS), respectively. Additionally, we compute the Word Error Rate (WER) evaluated via Whisper-large-v3 [42] to assess the accuracy of the spoken words.

Audio-visual alignment. We assess lip synchronization via SyncNet [9] and overall cross-modal semantic similarity using ImageBind [17]. We further introduce the Time Alignment Metric for Speech (TAMS) to quantify the synchronization between the synthesized speech and the designated temporal boundaries (*i.e.*, dialogue intervals).

Multi-shot consistency. We compute Subject Consistency (SC) and Background Consistency (BC) following VBench [25] to assess intra-shot temporal stability. Furthermore, we calculate Character and Background Inter-shot Consistency (CISC, BISC) scores utilizing ViCLIP [59] features. Finally, we introduce Shot Transition Accuracy (STA) to evaluate the model’s precision in executing shot transitions based on the narrative scripts.

5.3 Comparison with State-of-the-Art Methods

Baselines. We evaluate MAVIN against two primary paradigms of audio-visual generation on our 1K-sample high-fidelity benchmark. For a fair comparison, all models are evaluated without providing role references.

Cascaded generation models. We first evaluate multi-shot Text-to-Video (T2V) models (*i.e.*, VideoGen-of-Thought [79], MovieAgent [66], EchoShot [57], CineTrans [67], IC-LoRA [24], and StoryDiffusion [81]), followed by a Video-to-Audio (V2A) model (Hunyuan-Foley [46]) to generate the accompanying audio tracks. Notably, since IC-LoRA [24] and StoryDiffusion [81] only generate keyframes, we equip them with Wan2.2 [55] for generating the video frames.

Joint generation models. We compare MAVIN with end-to-end joint Text-to-Audio-Visual (T2AV) generation models (*i.e.*, JavisDiT [36], UniVerse-1 [56], OVI [39], and LTX-2 [18]), under the same inference settings as ours.

Quantitative results. As shown in Tab. 1, MAVIN outperforms all state-of-the-art baselines across 13 metrics. *(i) Cascaded generation models.* While these models [24, 57, 66, 67, 79, 81] achieve strong visual consistency (SC, BC) via dedicated T2V networks, their decoupled pipeline restricts the downstream V2A model to rigidly fit the generated videos. Since existing T2V models lack fine-grained lip motion, the downstream V2A models fail to generate intelligible speech, making WER and TAMS inapplicable. *(ii) Joint generation models.* Existing joint models [18, 36, 39, 56] improve basic audio-visual alignment but lack explicit mechanisms for hard cuts and long-term spatial correlations. Consequently, they suffer from character and background drifting (lower CISC/BISC) and fail to execute precise scene transitions ($STA < 0.60$). In contrast, our framework leverages boundary-aware attention to achieve a remarkable STA of 0.9897 and a CISC of 0.6319. Additionally, our framework achieves a state-of-the-art TAMS score of 0.8104, effectively ensuring that characters speak strictly within their designated temporal boundaries.

Qualitative results. In Fig. 4, we present qualitative comparisons. *(i) Cascaded generation models* generate visually plausible individual shots but suffer from severe cross-modal misalignment, including lip synchronization failures and abrupt audio discontinuities at transition boundaries. *(ii) Joint generation models* improve audio-visual coherence, but most of them still struggle with spatiotemporal consistency for multi-shot scenes, leading to noticeable character drifting and background inconsistency after cinematic cuts. Instead, our framework consistently preserves visual appearance and vocal timbre across shots with precise temporal alignment.

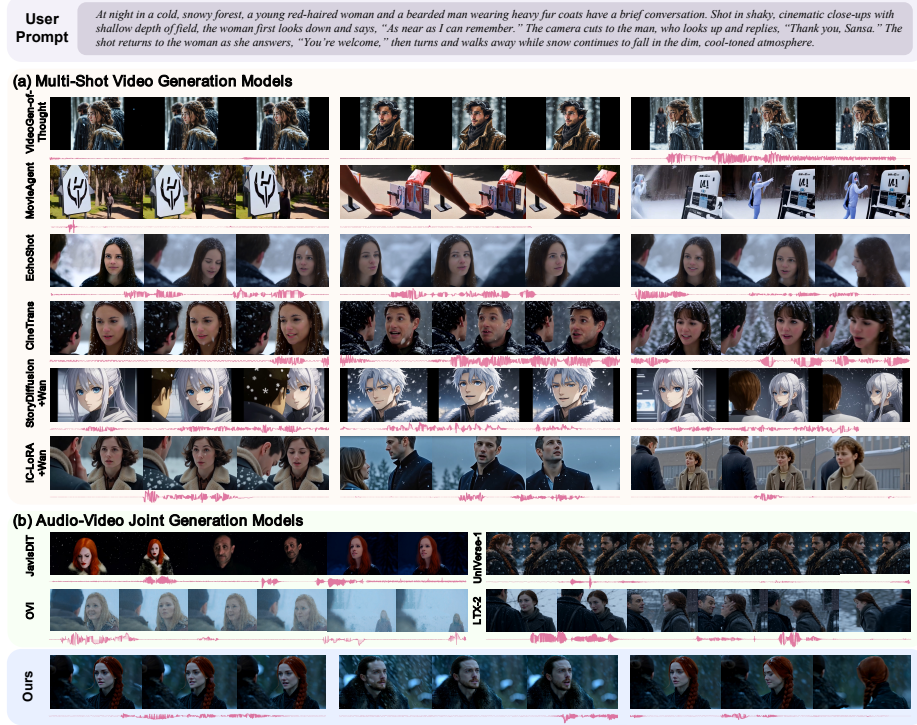


Fig. 4: Qualitative comparison with state-of-the-art relevant generation models.

5.4 Ablation Study

As presented in Tab. 2, we ablate key components of our framework to evaluate their individual contributions and report these quantitative results. Since some ablations target customized generation capabilities, we conduct these experiments with explicit video and audio references.

W/o BTR (Boundary-aware Token Routing). Removing BTR allows tokens to interact freely across designated temporal boundaries, leading to semantic leakage between shots. This uncontrolled interaction significantly degrades multi-shot consistency metrics (CISC, BISC, and STA).

W/o IM (Identity-aware Mask). Disabling IM allows identity anchors to interact freely across different shots and roles. This results in inter-subject crosstalk and identity blending in both visual appearance and vocal timbre, damaging character inter-shot consistency (CISC).

W/o PTE (Pretrained Timbre Encoder). Without the specialized PTE, the model struggles to decouple vocal timbre from other acoustic cues (*e.g.*, linguistic semantics, emotion, and prosody). This entanglement impacts the overall cross-modal synchronization (Sync, AV-IB).

Table 2: Quantitative ablation study. Removing modules consistently degrades quality.

Variant	Quality		Semantics			AV-Alignment			Multi-shot Consistency				
	FVD ↓	FAD ↓	TVS ↑	TAS ↑	WER ↓	Sync ↑	AV-IB ↑	TAMS ↑	SC ↑	BC ↑	CISC ↑	BISC ↑	STA ↑
w/o BTR	268.4	7.6	0.2310	0.2214	0.062	5.830	0.241	0.582	0.9631	0.9592	0.5520	0.7410	0.5580
w/o IM	256.9	7.3	0.2398	0.2301	0.055	5.970	0.252	0.787	0.9642	0.9617	0.5710	0.7680	0.8010
w/o PTE	242.5	7.0	0.2453	0.2368	0.051	6.010	0.258	0.792	0.9681	0.9643	0.6240	0.7920	0.9460
w/o MAP	261.1	7.5	0.2336	0.2248	0.058	5.910	0.244	0.569	0.9653	0.9618	0.5900	0.7700	0.6230
MAVIN	241.9	6.9	0.2463	0.2391	0.049	6.057	0.266	0.793	0.9682	0.9653	0.6397	0.7996	0.9886

W/o MAP (Multi-agent Scripting Pipeline). Directly feeding free-form prompts without MAP fails to provide the decoupled narrative semantics. Consequently, the model struggles with executing precise scene transitions (STA) and maintaining temporal alignment (TAMS).

5.5 User Study

Generation quality. To further assess subjective human preference, we conduct a user study across three key dimensions: *(i)* **Audio-Visual Quality (AVQ)**, evaluating the perceptual fidelity of generated videos and their accompanying audio; *(ii)* **Multi-shot Identity Consistency (MIC)**, measuring the stability of role attributes across shot boundaries without identity blending; and *(iii)* **Audio/Video Narrative Adherence (ANA/VNA)**, assessing the precise alignment between generated audio/video events and the structured scripting. In each experiment, participants are presented with generated audio-visual results from MAVIN and baseline methods, alongside the input free-form prompts. They are asked to select the result that best aligns with each evaluation dimension. We conduct the user study on Amazon Mechanical Turk (AMT) using 20 samples randomly selected from the MAVINSet dataset, with results polled from 25 volunteers. As shown in Tab. 3, our framework consistently achieves the highest preference scores across all dimensions, demonstrating superior human subjective performance.

Agent quality. To evaluate whether the multi-agent scripting pipeline effectively produces reliable hierarchical captions for downstream inference, we conduct an additional user study to measure the alignment between generated captions and intended narrative semantics. We ask 25 volunteers to evaluate 20 random samples and categorize the generated caption quality as “Failed”, “Borderline”, “Acceptable”, or “Perfect”. The resulting distribution is 1.2%, 4.2%, 10.4%, and 84.2%, respectively. With 94.6% of the ratings being “Acceptable” or higher, our scripting pipeline demonstrates strong robustness and reliability.

5.6 Applications

MAVIN supports several practical creative workflows for multi-shot audio-visual generation. Editable dialogue intervals in the shot-level script allow direct control over speaking pace. The generated storyboard can also be further modified

Table 3: Percentage (%) of user preference in the subjective evaluation.

<i>- Cascaded Generation Models: Multi-shot T2V + V2A</i>					<i>- Joint Generation Models: T2AV</i>				
Method	AVQ	MIC	ANA	VNA	Method	AVQ	MIC	ANA	VNA
VideoGen-of-Thought [79]	5.6	8.0	3.2	3.4	JavisDiT [36]	3.4	3.0	2.0	2.2
MovieAgent [66]	4.8	4.6	2.6	2.8	UniVerse-1 [56]	4.6	4.0	2.8	3.0
EchoShot [57]	8.2	7.0	5.2	5.4	OVI [39]	7.8	7.6	4.6	4.4
CineTrans [67]	9.4	6.6	5.0	5.2	LTX-2 [18]	6.6	7.2	2.4	2.4
IC-LoRA [24] + Wan [55]	6.6	4.8	2.8	3.2	Ours (MAVIN)	36.8	42.2	66.4	64.6
StoryDiff. [81] + Wan [55]	6.2	5.0	3.0	3.4					

for narrative editing, allowing story progression to be adjusted without redesigning the entire prompt. Moreover, assigning visual appearances and vocal timbres through image and audio references enables flexible identity customization. Finally, the multi-agent scripting pipeline can transform free-form user prompts into structured multi-shot audio-visual narratives, simplifying script planning for professional filmmaking workflows.

6 Conclusion

In this paper, we present MAVIN, the first framework for multi-shot audio-visual generation with customized narrative control. To resolve the challenge of temporal misalignment across complex timelines, we introduce boundary-aware attention, which enforces temporal alignment for shot transitions, dialogue intervals, and narrative events. Furthermore, we propose ID-aware propagation to overcome the limitations of multi-subject controllability, maintaining precise character consistency via image and audio references. To tackle incomplete scripting, we design a multi-agent scripting pipeline that translates free-form user inputs into hierarchical captions. Finally, we construct the MAVINSet dataset to facilitate robust training and evaluation. We believe our framework opens up a new avenue for integrating generative models into professional filmmaking workflows.

Limitation. Our model can produce videos of up to 15 seconds with a maximum of 3 customized characters. In its current state, deploying our framework for practical filmmaking workflows still requires iterative prompting and manual post-processing. For future exploration, by integrating stronger foundation models and scaling up the parameters, its practical utility could be further improved.

Acknowledgements

This work is supported by Beijing Major Science and Technology Project (Grant No. Z251100008125009) and National Natural Science Foundation of China (Grant No. 62136001).

References

1. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Asian Conference on Computer Vision (2020)
2. Cai, S., Ceylan, D., Gadelha, M., Huang, C.H.P., Wang, T.Y., Wetzstein, G.: Generative rendering: Controllable 4D-guided video generation with 2D diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
3. Chang, Z., Weng, S., Ouyang, H., Hong, Y., Lin, L., Li, S., Shi, B.: L-VOCAL: Language-based video colorization with audio alignment. *International Journal of Computer Vision* **134**(5), 208 (2026)
4. Chatterjee, M., Cherian, A.: Sound2Sight: Generating visual dynamics from sound and context. In: European Conference on Computer Vision (2020)
5. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: VGGSound: A large-scale audio-visual dataset. *IEEE International Conference on Acoustics, Speech and Signal Processing* (2020)
6. Chen, K., Li, X., Li, Y., Tong, Y., Wu, J., Zeng, Y., Zhang, J., Zhou, Q.: MotionBooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems* (2024)
7. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
8. 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)
9. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Asian Conference on Computer Vision (2016)
10. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025)
11. ElevenLabs: Introducing Scribe v2. <https://elevenlabs.io/blog/introducing-scribe-v2> (2026), accessed: 2026-03-05
12. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: CLAP: learning audio concepts from natural language supervision. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2023)
13. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113* (2025)
14. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: *IEEE international conference on acoustics, speech and signal processing* (2017)
15. Geng, D., Herrmann, C., Hur, J., Cole, F., Zhang, S., Pfaff, T., Lopez-Guevara, T., Aytar, Y., Rubinstein, M., Sun, C., Wang, O., Owens, A., Sun, D.: Motion Prompting: Controlling video generation with motion trajectories. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025)

16. Ghermi, R., Wang, X., Kalogeiton, V., Laptev, I.: Long Story Short: Story-level video understanding from 20k short films. arXiv preprint arXiv:2406.10221 (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: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
18. 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)
19. Haji-Ali, M., Menapace, W., Siarohin, A., Skorokhodov, I., Canberk, A., Lee, K.S., Ordonez, V., Tulyakov, S.: AV-Link: Temporally-aligned diffusion features for cross-modal audio-video generation. In: International Conference on Computer Vision (2025)
20. He, J., Liu, H., Li, J., Huang, Z., Yu, Q., Ouyang, W., Liu, Z.: Cut2next: Generating next shot via in-context tuning. In: ACM SIGGRAPH Asia Conference Papers (2025)
21. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
22. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: CogVideo: Large-scale pretraining for text-to-video generation via transformers. In: International Conference on Learning Representations (2023)
23. Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., et al.: Qwen3-TTS technical report. arXiv preprint arXiv:2601.15621 (2026)
24. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)
25. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: VBench: Comprehensive benchmark suite for video generative models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
26. Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al.: NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. arXiv preprint arXiv:2403.03100 (2024)
27. Kara, O., Singh, K.K., Liu, F., Ceylan, D., Rehg, J.M., Hinz, T.: Shotadapter: Text-to-multi-shot video generation with diffusion models. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
28. 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)
29. Kushwaha, S.S., Tian, Y.: VinTAGe: Joint video and text conditioning for holistic audio generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
30. Lei, G., Wang, C., Zhang, R., Wang, Y., Li, H., Xu, W.: AnimateAnything: Consistent and controllable animation for video generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
31. Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., Zhu, S.: Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)

32. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: MagicMotion: Controllable video generation with dense-to-sparse trajectory guidance. In: International Conference on Computer Vision (2025)
33. Li, Z., Rahmani, H., Ke, Q., Liu, J.: LongDiff: Training-free long video generation in one go. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
34. Lin, H., Zala, A., Cho, J., Bansal, M.: VideoDirectorGPT: Consistent multi-scene video generation via llm-guided planning. In: Conference on Language Modeling (2024)
35. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
36. Liu, K., Li, W., Chen, L., Wu, S., Zheng, Y., Ji, J., Zhou, F., Luo, J., Liu, Z., Fei, H., Chua, T.S.: JarvisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. arXiv preprint arXiv:2503.23377 (2025)
37. Liu, K., Zheng, Y., Wang, K., Wu, S., Zhang, R., Luo, J., Hatzinakos, D., Liu, Z., Fei, H., Chua, T.S.: JarvisDiT++: Unified modeling and optimization for joint audio-video generation. arXiv preprint arXiv:2602.19163 (2026)
38. Long, F., Qiu, Z., Yao, T., Mei, T.: VideoStudio: Generating consistent-content and multi-scene videos. European Conference on Computer Vision (2024)
39. Low, C., Wang, W., Katyal, C.: Ovi: Twin backbone cross-modal fusion for audio-video generation. arXiv (2025)
40. Luo, S., Yan, C., Hu, C., Zhao, H.: Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. Advances in Neural Information Processing Systems (2023)
41. Montanaro, A., Savant Aira, L., Aiello, E., Valsesia, D., Magli, E.: Motioncraft: Physics-based zero-shot video generation. Advances in Neural Information Processing Systems (2024)
42. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning (2023)
43. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
44. 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. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
45. Schuhmann, C.: Improved aesthetic predictor. <https://github.com/christophschuhmann/improved-aesthetic-predictor> (2022)
46. Shan, S., Li, Q., Cui, Y., Yang, M., Wang, Y., Yang, Q., Zhou, J., Zhong, Z.: HunyuanVideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. arXiv preprint arXiv:2508.16930 (2025)
47. Shi, B., Tjandra, A., Hoffman, J., Wang, H., Wu, Y.C., Gao, L., Richter, J., Le, M., Vyas, A., Chen, S., et al.: SAM Audio: Segment anything in audio. arXiv preprint arXiv:2512.18099 (2025)
48. Soucek, T., Lokoc, J.: TransNet V2: An effective deep network architecture for fast shot transition detection. In: Cai, J., Kankanhalli, M.S., Prabhakaran, B., Boll, S., Subramanian, R., Zheng, L., Singh, V.K., César, P., Xie, L., Xu, D. (eds.) ACM International Conference on Multimedia (2024)

49. 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. *ACM International Conference on Multimedia* (2024)
50. Tan, Z., Yang, X., Liu, S., Wang, X.: Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260* (2024)
51. Team, O., Yu, D., Chen, M., Chen, Q., Luo, Q., Wu, Q., Cheng, Q., Li, R., Liang, T., Zhang, W., et al.: MOVA: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794* (2026)
52. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: *European Conference on Computer Vision* (2020)
53. Tjandra, A., Wu, Y.C., Guo, B., Hoffman, J., Ellis, B., Vyas, A., Shi, B., Chen, S., Le, M., Zacharov, N., Wood, C., Lee, A., Hsu, W.N.: Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139* (2025)
54. 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)
55. 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)
56. Wang, D., Zuo, W., Li, A., Chen, L.H., Liao, X., Zhou, D., Yin, Z., Dai, X., Jiang, D., Yu, G.: Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155* (2025)
57. Wang, J., Sheng, H., Cai, S., Zhang, W., Yan, C., Feng, Y., Deng, B., Ye, J.: EchoShot: Multi-shot portrait video generation. In: *Advances in Neural Information Processing Systems* (2025)
58. Wang, K., Deng, S., Shi, J., Hatzinakos, D., Tian, Y.: Av-dit: Taming image diffusion transformers for efficient joint audio and video generation. *ACM International Conference on Multimedia* (2025)
59. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: InternVid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942* (2023)
60. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: MotionCtrl: A unified and flexible motion controller for video generation. *ACM SIGGRAPH Conference Papers* (2024)
61. Weng, S., Zheng, H., Chang, Z., Li, S., Shi, B., Wang, X.: Audio-sync video generation with multi-stream temporal control. In: *Advances in Neural Information Processing Systems* (2025)
62. Wu, B., Zou, C., Li, C., Huang, D., Yang, F., Tan, H., Peng, J., Wu, J., Xiong, J., Jiang, J., et al.: HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870* (2025)
63. 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>
64. Wu, H., Chen, K., Luo, Y., Qiao, R., Ren, B., Liu, H., Xie, W., Shen, L.: Scene consistency representation learning for video scene segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022)

65. Wu, W., Liu, M., Zhu, Z., Xia, X., Feng, H., Wang, W., Lin, K.Q., Shen, C., Shou, M.Z.: Moviebench: A hierarchical movie level dataset for long video generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
66. Wu, W., Zhu, Z., Shou, M.Z.: Automated movie generation via multi-agent CoT planning. arXiv preprint arXiv:2503.07314 (2025)
67. Wu, X., Gao, B., Qiao, Y., Wang, Y., Chen, X.: CineTrans: Learning to generate videos with cinematic transitions via masked diffusion models. arXiv preprint arXiv:2508.11484 (2025)
68. Wu, Z., Siarohin, A., Menapace, W., Skorokhodov, I., Fang, Y., Chordia, V., Gilitschenski, I., Tulyakov, S.: Mind the time: Temporally-controlled multi-event video generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
69. 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 (2024)
70. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., et al.: Qwen3-Omni technical report. arXiv preprint arXiv:2509.17765 (2025)
71. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., et al.: VLIPP: Towards physically plausible video generation with vision and language informed physical prior. In: International Conference on Computer Vision (2025)
72. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations (2025)
73. Zhang, P., Jia, Z., Liu, K., Weng, S., Li, S., Shi, B.: STAGE: Storyboard-anchored generation for cinematic multi-shot narrative. arXiv preprint arXiv:2512.12372 (2025)
74. Zhao, C., Liu, M., Wang, W., Chen, W., Wang, F., Chen, H., Zhang, B., Shen, C.: MovieDreamer: Hierarchical generation for coherent long visual sequence. arXiv preprint arXiv:2407.16655 (2024)
75. Zhao, L., Feng, L., Ge, D., Yi, F., Zhang, C., Zhang, X.L., Li, X.: Uniform: A unified diffusion transformer for audio-video generation. arXiv e-prints (2025)
76. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: ControlVideo: conditional control for one-shot text-driven video editing and beyond. Science China Information Sciences (2025)
77. Zheng, H., Weng, S., Liu, J., Yang, S., Shi, B., Wang, X.: Audio-sync video instance editing with granularity-aware mask refiner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23150–23160 (2026)
78. Zheng, H., Yang, Y., Yang, S., Weng, S., Shi, B.: InstructAV2AV: Instruction-guided audio-video joint editing. arXiv preprint arXiv:2605.18467 (2026)
79. Zheng, M., Xu, Y., Huang, H., Ma, X., Liu, Y., Shu, W., Pang, Y., Tang, F., Chen, Q., Yang, H., et al.: VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention. arXiv preprint arXiv:2412.02259 (2024)
80. Zhou, Y., Wang, Z., Fang, C., Bui, T., Berg, T.L.: Visual to sound: Generating natural sound for videos in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
81. Zhou, Y., Zhou, D., Cheng, M.M., Feng, J., Hou, Q.: StoryDiffusion: Consistent self-attention for long-range image and video generation. Advances in Neural Information Processing Systems (2024)