# DÉCOUPAGE: DIRECTORIAL AUDIO-VISUAL CONTROL FOR LONG MULTI-SHOT VIDEO GENERATION

**Anonymous authors**
Paper under double-blind review

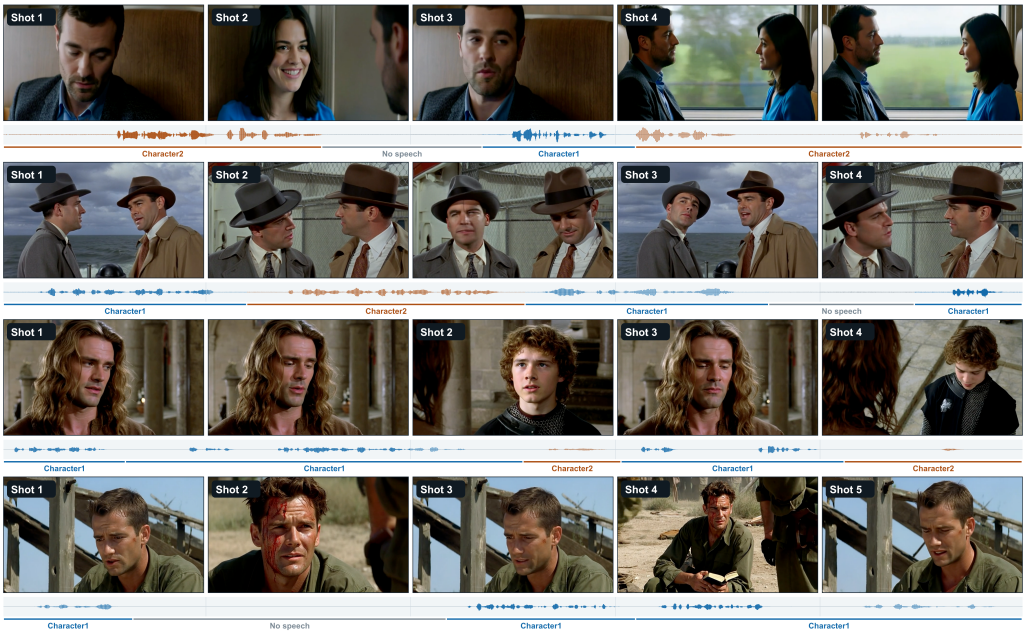


Figure 1: **Screenplay-directed audio-video generation.** DÉCOUPAGE directs visual and audio timing tracks across multi-shot scenes. Four 15-second examples demonstrate precise control over speech bridging across cuts (J-/L-cuts) and silent listener reactions, with all shot transitions and synthesized audio strictly following the screenplay timeline.

## ABSTRACT

In cinematic storytelling, sound and visual transitions do not always coincide in lockstep: dialogue frequently bridges across cuts to reveal reactions or introduces new scenes via J- and L-cuts, while deliberate silences pace the scene. Existing audio-visual generative frameworks, however, lack explicit mechanisms to coordinate independent visual and acoustic timing over extended horizons. As narrative duration and shot complexity scale, these systems struggle to maintain long-range dialogue pacing—often hallucinating speech during intended silences or erroneously driving mouth movements on silent listeners. To enable fine-grained cinematic control, we present DÉCOUPAGE, a screenplay-driven framework for long-form, multi-shot, and multi-character joint audio-video generation. Rather than coupling modality boundaries by default, DÉCOUPAGE uncouples generation into independently bounded visual and acoustic tracks governed by persistent character identities. DÉCOUPAGE builds on a joint audio-video diffusion model with dual-schedule temporal cross-attention, incorporating a directorial Lip-sync mechanism that selectively decouples speech conditioning from visual facial dynamics during off-screen dialogue. On multi-shot conversational benchmarks, DÉCOUPAGE achieves precise control over shot transitions, audio timing, and visibility-aware lip synchronization, directing 15- to 60-second screenplay-driven audio-video generation with multimodal coherence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

# 1 INTRODUCTION

Cinematic storytelling relies fundamentally on the freedom to uncouple what is seen from what is heard. A camera cut marks a discrete visual shift, whereas dialogue routinely spills across visual boundaries: in J- and L-cuts, speech anticipates an incoming scene or carries past a cut to capture a listener's silent reaction. These expressive devices require independently scheduled visual and acoustic timelines, balanced by deliberate silences and tightly synchronized articulation whenever a visible character actually speaks.

Recent joint audio-video foundation models have enabled impressive synchronized multimodal synthesis (HaCohen et al., 2026; Liu et al., 2024; Zhang et al., 2025; Ma et al., 2026). However, they inherently inherit a *rigid synchrony assumption*—a built-in tendency to co-terminate shot boundaries, acoustic events, and facial dynamics within tightly coupled windows. In multi-shot conversations, this coupling falls short: dialogue easily drifts across cuts, scripted silences leak spurious speech, and off-screen voices erroneously trigger mouth movements on silent listeners. These failures expose two distinct directorial challenges: acoustic continuity and deliberate silences must be maintained independently across camera cuts, while active speech must animate only visible speakers. Crucially, extending visual continuity alone (Meng et al., 2025; 2026) leaves these cross-modal pacing conflicts unresolved.

To address this gap, we introduce DÉCOUPAGE, a screenplay-driven framework that brings fine-grained directorial control to multi-shot, multi-character audio-video generation. Rather than binding modality boundaries together, DÉCOUPAGE decomposes generation into two independently timed visual and acoustic tracks aligned by a shared clock and persistent character identities. Under this formulation, dialogue can naturally span visual transitions while speakers move on- and off-screen. To support this decoupled design, we construct continuous 15- and 60-second training sequences by merging adjacent cinematic shots, pairing Whisper-derived speech timelines with Gemini-annotated character roles (Radford et al., 2023; Klein, 2023; Reid et al., 2024) to natively retain sound bridges and reaction shots.

Building upon a joint diffusion backbone adapted via LoRA (Hu et al., 2021), DÉCOUPAGE introduces two complementary mechanisms for decoupled control. First, dual-schedule temporal cross-attention enforces shot-aware video–text intervals and timestamp-aware audio–text intervals during denoising, steering multi-shot visual composition and acoustic dialogue schedules along independent timelines. Second, to resolve listener attribution errors, we introduce the directorial Lip-sync mechanism, an inference-time method that selectively regulates speech-driven articulation based on character visibility. During inference, whenever an utterance occurs off-screen, it selectively diverts the video stream's audio cross-attention to an auxiliary non-speech trajectory. This bypasses unwanted speech-driven facial articulation on reaction shots while retaining the main joint audio-generation path without requiring additional learned parameters.

Evaluating cinematic multi-shot conversations demands assessing both temporal fidelity and selective synchronization. Beyond standard generation metrics, a valid framework must verify that dialogue stays within its scripted window, that deliberate silences remain quiet, and that visible listeners stay quiescent during off-screen speech. Across 15- and 60-second conversational benchmarks, DÉCOUPAGE substantially outperforms existing joint generators in shot-cut adherence, turn containment, and silence preservation, while effectively eliminating spurious lip synchronization on reaction shots.

Our contributions are threefold:

- **Screenplay Formulation and Data Pipeline** We formulate multi-shot generation through independently bounded visual and acoustic tracks, supported by an automated pipeline that curates continuity-preserving 15- and 60-second scenes with persistent character identities.
- **Directorial Audio-Visual Generation** We employ dual-schedule temporal cross-attention for independently timed shot and acoustic control, with a training-free directorial Lip-sync mechanism that reduces unintended listener articulation during off-screen dialogue.
- **Cinematic Benchmark and Scalability** We establish a screenplay-aligned benchmark with targeted metrics for cut fidelity, silence preservation, and listener quiescence. DÉCOUPAGE achieves state-of-the-art script adherence on 15-second scenes and robustly sustains these directorial controls when scaled to minute-long narratives.

# 2 Related Work

## 2.1 Long-form and multi-shot video generation

Long-form generation has increasingly separated narrative planning from visual rendering. VideoDirectorGPT translates a prompt into an explicit scene plan before generation (Lin et al., 2024), MovieDreamer combines a hierarchical autoregressive narrative model with diffusion rendering (Zhao et al., 2025), and StoryDiffusion introduces cross-frame consistency mechanisms for image and video stories (Zhou et al., 2024). VideoAuteur further uses rolling visual context and keyframe-based rendering to produce long narrative videos (Xiao et al., 2025). These methods establish the value of high-level story structure, but their plans do not directly specify an editable soundtrack.

Another line makes the shot itself an explicit unit of control. ShotAdapter models transitions with dedicated tokens and local attention masks (Kara et al., 2025); HoloCine combines shot-local text windows with sparse inter-shot attention for minute-scale generation (Meng et al., 2025); CausalCine retrieves relevant history during autoregressive shot progression (Meng et al., 2026); and ShotDirector controls camera parameters and editing patterns with shot-aware conditioning (Wu et al., 2025). These works primarily organize visual continuity; they do not directly model the independently timed flow of dialogue across cuts. We instead treat the cut schedule and the acoustic-event schedule as independently timed but jointly generated tracks, so dialogue and sound can begin, end, or continue across a visual edit.

## 2.2 Joint audio-video generation

Joint audio-video models replace post-hoc sound generation with coupled denoising. SyncFlow uses interacting audio and video diffusion transformers for text-conditioned synthesis (Liu et al., 2024); JavisDiT adds hierarchical spatiotemporal priors for synchronized generation (Liu et al., 2026a); and Animate and Sound an Image jointly expands a still image into motion and audio (Wang et al., 2025). More recent systems develop scalable dual-stream architectures and stronger cross-modal interaction, including LTX-2 (HaCohen et al., 2026), UniAVGen (Zhang et al., 2025), Cross-Modal Context Learning (Ma et al., 2026), and Harmony (Hu et al., 2026).

Closed-source systems also support native joint generation. Seedance 1.5 Pro uses a dual-branch diffusion transformer with cross-modal interaction (Esser et al., 2024), while Seedance 2.0 extends audio-video generation to multimodal reference and editing inputs (Team Seedance et al., 2025; 2026). Veo 3 jointly denoises temporal audio latents and spatiotemporal video latents (Google DeepMind, 2025). Kling-Omni's technical report describes unified multimodal video generation and editing (Kling Team et al., 2025); the subsequent Kling 3.0 release adds native audio, multi-character dialogue, and shot-level storyboard control (Kuaishou Technology, 2026). These capabilities make explicit shot and dialogue control a relevant comparison axis, beyond perceptual quality alone.

Controllable variants begin to expose speech and cinematic structure. MM-Sonate adds phoneme-level control and reference-timbre conditioning for intelligible, identity-preserving speech (Qiang et al., 2026), while CineDance and MAVIN extend joint generation toward multi-shot narratives with structured annotations or hierarchical scripts (Chen et al., 2026; Liu et al., 2026b). These advances emphasize synchronized audiovisual coherence, but symmetric cross-modal interaction alone does not specify when speech should animate a visible face. In particular, an off-screen utterance and a listener's reaction are temporally related without requiring lip synchronization. Our rigid synchrony critique concerns this control gap, rather than an architectural prohibition on off-screen sound. DÉCOUPAGE makes shot and acoustic schedules independently addressable, and selectively regulates speech-driven articulation based on shot-level character visibility. We evaluate both temporal alignment and visibility-conditioned coupling explicitly.

## 2.3 Multi-Character Speech and Cinematic Articulation

Audio-driven talking-face generation has evolved from single-speaker lip synchronization (Prajwal et al., 2020; Zhang et al., 2024; Li et al., 2024) to multi-character identity routing through spatial-temporal masks and label-aware positional embeddings (Kong et al., 2025; Huang et al., 2025).

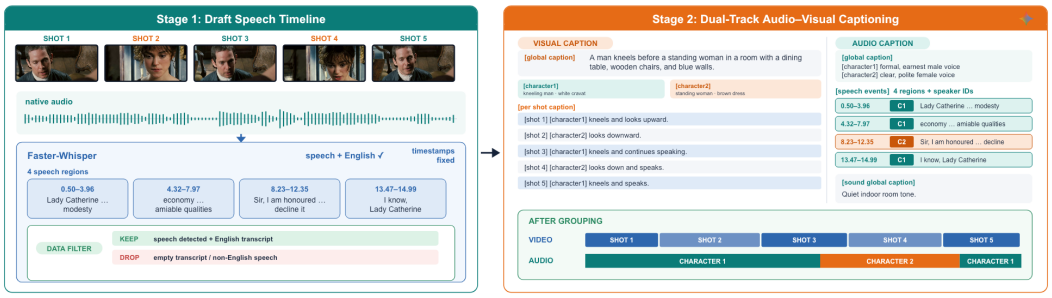


Figure 2: Joint audio-video data pipeline. Adjacent shots form audiovisual sequences. Faster-Whisper establishes a fixed speech timeline and filters out empty or non-English transcripts. Gemini 3.0 Pro then produces visual and audio captions with persistent character identities. Finally, adjacent speech events assigned to the same character are grouped into character-level audio intervals.

These methods typically animate visible target faces from a supplied audio signal. Their central task is to produce the intended articulation, whereas cinematic dialogue also requires deciding when audible speech should leave visible listeners unarticulated. Reaction shots and J- and L-cuts make this distinction explicit: a speaker's voice can continue after their face has left the frame. Conversational diagnostics likewise highlight the importance of speaker–listener attribution (Zhou et al., 2026). Our focus is this role decision within native joint generation. We leverage shot-level visibility to dynamically regulate the audio-to-video update, complementing asynchronous dialogue scheduling without requiring a separate talking-face synthesis stage.

## 3 METHOD

We define a structured screenplay as two independently timed tracks specifying visual shots and acoustic events, with shared character identities and a common scene clock. Given this screenplay, DÉCOUPAGE jointly generates a multi-shot video and its soundtrack. To follow these independent schedules, we applies dual-schedule temporal cross-attention to the video and audio streams, allowing dialogue to span visual cuts. A training-free directorial Lip-sync mechanism further regulates audio-to-video cross-attention residuals based on shot-level speaker visibility, reducing unintended listener articulation during off-screen dialogue.

### 3.1 SCREENPLAY FORMULATION AND DATA PIPELINE

**Dual-track representation.** We represent the screenplay as a visual track $\mathcal{S}_V = \{(I_i^V, c_i^V, \mathcal{C}_i)\}_{i=1}^{N_V}$ and an acoustic track $\mathcal{S}_A = \{(I_j^A, c_j^A, \rho_j)\}_{j=1}^{N_A}$. Here, $I_i^V$ is a shot interval, $c_i^V$ describes its framing and action, and $\mathcal{C}_i$ contains its visible character identities. Each acoustic event has an interval $I_j^A$ and a description $c_j^A$ specifying dialogue or non-speech sound; $\rho_j$ identifies the speaker when the event contains speech. Character identities persist across shots and dialogue turns.

Both tracks use the same continuous scene clock $t \in [0, T]$, where $T$ is the sequence duration, but their boundaries are independent. In particular, an acoustic event may overlap multiple shots. Its speaker identity remains fixed, while its on-screen or off-screen state changes according to the visible characters in each overlapping shot. J- and L-cuts are thus specified by placing acoustic events relative to visual cuts.

**Continuity-preserving sequence construction.** Following HoloCine's shot-based preprocessing (Meng et al., 2025), we merge temporally adjacent shots from the same source into 15- and 60-second sequences. Video and native audio are resampled together, preserving their alignment and dialogue across cuts. We remove duplicate sequence plans and retain the shot boundaries.

**Data captioning.** We use Faster-Whisper (Radford et al., 2023; Klein, 2023) to obtain transcripts and speech timestamps, which remain fixed during annotation. Given cut-aware frame samples, the assembled audio, and timestamped transcripts, Gemini 3.0 Pro produces visual and acoustic tracks with shared character identities. The visual track describes each shot's framing, visible characters, and actions. The acoustic track annotates dialogue and speaker identities along the Whisper timeline,

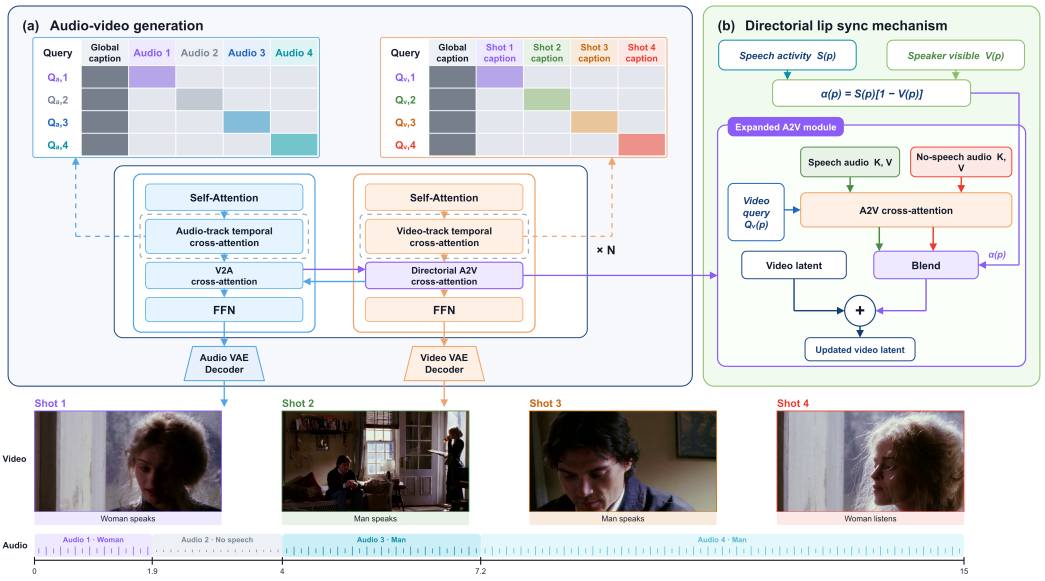


Figure 3: Overview of DÉCOUPAGE. (a) Dual-track temporal cross-attention guides joint audio-video generation along independent visual and acoustic schedules, allowing dialogue to continue across shot boundaries. (b) The directorial lip-sync mechanism blends speech and no-speech A2V residuals based on speaker visibility to suppress unintended listener articulation.

marks intervals without speech as `<NO SPEECH>`, and uses a `[sound global caption]` to describe foley, ambient sounds, background music, and the scene's overall acoustic atmosphere.

Recurring characters receive persistent `[characterX]` identifiers across shots and speech events, including off-screen dialogue; voices whose speakers never appear in the sequence are labeled `[bg]`. We group speech annotations by character and exclude audio annotations involving more than four distinct speaking characters. We retain reaction shots and off-screen dialogue rather than applying a global lip-sync filter that could exclude these valid examples.

### 3.2 DUAL-SCHEDULE TEMPORAL CROSS-ATTENTION

To coordinate multi-shot visuals and continuous audio without enforcing rigid modality synchrony, we implement dual-schedule temporal cross-attention. Our dual-schedule isolates the visual and acoustic text conditions along their respective screenplay timelines: a visual cut shifts the accessible shot descriptions for video queries, while an ongoing speech or acoustic event conditioning remains intact across that transition.

Specifically, for each modality $X \in \{V, A\}$, let $\tau_X(p)$ denote the scene timestamp corresponding to latent query token $p$. The textual conditioning keys are partitioned into global context tokens $\mathcal{G}$, shared register tokens $\mathcal{R}$, and interval-specific text tokens $\mathcal{T}_i^X$ associated with interval $I_i^X$. Notably, $\mathcal{G} = \mathcal{G}_V \cup \mathcal{G}_A$ simultaneously encapsulates sequence-level global descriptions for both video and audio. The accessible key index set $\mathcal{A}_X(p)$ for query $p$ is formally defined as:

$$\mathcal{A}_X(p) = \mathcal{G} \cup \mathcal{R} \cup \bigcup_{i:\, \tau_X(p) \in I_i^X} \mathcal{T}_i^X. \tag{1}$$

For the visual track ($X = V$), each query token $p$ strictly attends to its current shot's prompt $I_i^V$. For the acoustic track ($X = A$), the union naturally accommodates multiple concurrent acoustic descriptions active at timestamp $\tau_A(p)$, seamlessly allowing ambient foley and environmental sounds.

We enforce this dual-schedule temporal cross-attention via an additive attention mask $M_{XT} \in \mathbb{R}^{N_X \times L_T}$:

$$M_{XT}[p, k] = \begin{cases} 0, & k \in \mathcal{A}_X(p), \\ -\kappa, & \text{otherwise}, \end{cases} \tag{2}$$

where $\kappa$ is a large positive scalar ($e.g., 10^4$). The routed cross-attention is computed as:

$$\mathrm{Attn}_{XT} = \mathrm{softmax}\left(\frac{Q_X K_{T,X}^\top}{\sqrt{d_X}} + M_{XT}\right) V_{T,X}, \tag{3}$$

where $Q_X$ denotes the latent queries of modality $X$, and $K_{T,X}, V_{T,X}$ are the text conditioning keys and values. Because recurring characters share consistent identifiers (***e.g.***, `[character1]`) across different intervals $\mathcal{T}_i^A$ and $\mathcal{T}_i^V$, this dual-schedule formulation grounds long-range identity consistency across alternating visual shots and asynchronous dialogue turns.

### 3.3 Directorial Lip-Sync Mechanism

While dual-schedule temporal cross-attention specifies *when* camera cuts occur and *when* dialogue begins or ends, it does not dictate *whether* active speech should animate a visible character. During listener/reaction shots, dialogue naturally carries over while the camera captures a silent listener. Rather than leaving the video features unconstrained without acoustic guidance, we supply an explicit quiescent signal by deriving an auxiliary audio trajectory conditioned on `<NO SPEECH>`. This training-free mechanism selectively suppresses speech-driven lip motion during off-screen dialogue, while preserving the listener's natural facial expressions and emotional reactions undisturbed.

**Paired speech and no-speech trajectories.** At each denoising step, we obtain audio states from the native joint trajectory and an auxiliary, paired-noise audio-only trajectory conditioned on `<NO SPEECH>`. At block $l$, these trajectories supply two sets of audio keys and values, $(K^{(l)}_{A,\mathrm{native}}, V^{(l)}_{A,\mathrm{native}})$ and $(K^{(l)}_{A,\mathrm{ns}}, V^{(l)}_{A,\mathrm{ns}})$. Both are evaluated against the same native video queries:

$$r_b^{(l)} = \mathcal{F}^{(l)}_{\mathrm{A2V}}\left(Q_V^{(l)}, K_{A,b}^{(l)}, V_{A,b}^{(l)}\right), \quad b \in \{\mathrm{native}, \mathrm{ns}\}, \tag{4}$$

where $\mathcal{F}^{(l)}_{\mathrm{A2V}}$ denotes the block's audio-to-video cross attention and residual-output mapping. The alternative residual therefore uses audio features from the corresponding block and denoising step, rather than a zeroed audio input or a separate full video prediction.

**Directorial A2V residual blending.** We blend the paired audio to video (A2V) residuals according to whether the screenplay places the active speaker on screen. For a video query $p$ in shot $i$, let $S(p) \in \{0, 1\}$ indicate scheduled speech activity, and let $V(p) = \mathbb{I}(\rho_j \in \mathcal{C}_i)$ indicate whether the active speaker $\rho_j$ belongs to the set of characters $\mathcal{C}_i$ shown in that shot. These screenplay directives determine the blending weight

$$\alpha(p) = S(p)\big(1 - V(p)\big), \tag{5}$$

which we apply to the paired residuals at each block $l$:

$$\tilde{r}^{(l)}_{\mathrm{A2V}}(p) = \big(1 - \alpha(p)\big) r^{(l)}_{\mathrm{native}}(p) + \alpha(p) r^{(l)}_{\mathrm{ns}}(p). \tag{6}$$

During off-screen dialogue ($S = 1, V = 0$), $\alpha = 1$ replaces the native A2V residual with its no-speech counterpart, providing quiescent acoustic guidance for the visible listener. When the speaker is on screen ($V = 1$) or no speech is scheduled ($S = 0$), $\alpha = 0$ retains the native residual. The control is applied uniformly across video queries within each temporal intersection of a shot and a speech interval, requiring no spatial face masks. At the A2V residual addition, the video state is updated as

$$h^{(l)}_{V,\mathrm{out}}(p) = h^{(l)}_{V,\mathrm{base}}(p) + \tilde{r}^{(l)}_{\mathrm{A2V}}(p). \tag{7}$$

where $h^{(l)}_{V,\mathrm{base}}$ is the video state before this residual is added. Subsequent block operations proceed normally. This intervention changes the audio-derived update received by the visible scene during off-screen speech, while retaining native coupling for visible speakers.

The auxiliary audio trajectory requires no video backbone computation or video-to-audio (V2A) cross-attention, functioning solely to supply the no-speech audio states for the alternative A2V residual. Meanwhile, the main generation branch retains its native bidirectional coupling without directly replacing or blending the main branch's audio prediction. This lightweight design introduces minimal computational overhead, as verified by the per-step runtime in Section 4.5.

| Model | Video Quality | | | Audio Quality | | | | Voice Identity | | AV Align. |
|---|---|---|---|---|---|---|---|---|---|---|
| | Cut F1 ↑ | Cnt MAE ↓ | Imaging ↑ | WER ↓ | Cont. ↑ | Leak ↓ | PQ ↑ | Consist. ↑ | Margin ↑ | Sync-C ↑ |
| **15-second Evaluation Dataset** | | | | | | | | | | |
| LTX-2.3 Base | 0.26 | 1.23 | 0.68 | 35.69 | 52.68 | 18.69 | 6.18 | 0.42 | 0.12 | 2.52 |
| MiniMax-H3 | 0.29 | 0.45 | **0.70** | 10.57 | 41.95 | 27.14 | **7.27** | 0.50 | 0.14 | **5.00** |
| MOVA (I2VA) | 0.18 | 7.38 | 0.56 | 38.22 | 39.33 | 28.47 | 6.26 | 0.36 | 0.01 | 3.24 |
| **Ours** | **0.99** | **0.05** | 0.68 | **4.59** | **93.01** | **1.12** | 6.91 | **0.53** | **0.29** | 3.59 |
| **60-second Evaluation Dataset** | | | | | | | | | | |
| LTX-2.3 Base | 0.05 | 5.27 | 0.61 | 43.68 | 11.88 | 32.71 | 6.26 | 0.42 | 0.06 | 2.77 |
| **Ours** | **0.76** | **1.81** | **0.66** | **7.13** | **91.12** | **4.46** | **6.84** | **0.57** | **0.22** | **4.00** |

Table 1: **Quantitative results on the 15-second and 60-second benchmarks.** Metrics WER, Cont., and Leak are reported in percentage (%). Bold font indicates the best performance within each duration group.

# 4 EXPERIMENTS

## 4.1 DATASETS AND EXPERIMENTAL SETUP

**Datasets and Evaluation Benchmark.** The automated curation pipeline (Section 3.1) yields approximately 30,000 cleaned, temporally aligned video–audio sequences for model adaptation. To benchmark performance under complex conversational dynamics, we construct an evaluation suite comprising 200 15-second and 200 60-second multi-character screenplays. Each screenplay explicitly specifies shot boundaries, recurring character identities, dialogue lines, and fine-grained acoustic timestamps, enabling standardized evaluation across both short- and long-horizon video generation.

**Implementation Details and Baselines.** We instantiate our approach upon the LTX-2.3 Base joint audio–video backbone (HaCohen et al., 2026), fine-tuning with LoRA (rank $r = 32$, initial learning rate $5 \times 10^{-5}$ with a linear decay schedule) across 32 NVIDIA H20 GPUs. We compare DÉCOUPAGE against representative open and proprietary joint audiovisual generation models, including the base model LTX2.3, MiniMax-H3 (MiniMax, 2026), and MOVA (SII-OpenMOSS Team et al., 2026) in its image-to-video-audio (I2VA) configuration. All baselines are supplied with identical screenplay directives and evaluated using their official inference and sampling settings.

## 4.2 EVALUATION METRICS

**Cinematic Shot Control.** We detect scene transitions at a fixed threshold of 0.18 and match them one-to-one to screenplay cuts within 0.5 seconds. We report **Cut F1** and the mean absolute error in cut count (**Count MAE**). In addition, we evaluate the **Imaging Quality** dimension of VBench (Huang et al., 2023) to assess overall video fidelity.

**Dialogue and Pacing Control.** We transcribe generated audio with Faster-Whisper (Radford et al., 2023; Klein, 2023) and compute corpus-level **word error rate (WER)** against the reference dialogue. **Containment** (↑) is the percentage of target turns whose aligned speech overflows the scripted window by no more than 0.2 seconds for the 15-second evaluation dataset or 0.4 seconds for the 60-second evaluation dataset. **Silence Leakage** (↓) is the duration-weighted overlap between recognized words and the interior of scripted silence intervals. We assess acoustic naturalness with **AudioBox P-Q** (Vyas et al., 2023). Exact definitions and support rules appear in Appendix A.2.

**Vocal and Audiovisual Consistency.** We evaluate character voice identity with ECAPA-TDNN embeddings (Desplanques et al., 2020). **Role Consistency** is the cosine similarity of same-character voice embeddings across utterances; **Role Margin** is its within-video margin over different-character similarity. **Sync-C** is the SyncNet confidence (Chung & Zisserman, 2016), which measures audiovisual alignment between a visible person's lip movements and the speech audio. To detect whether non-speaking characters are erroneously driven by off-screen dialogue, we introduce **TemporalBias**, which measures spurious audio-visual coupling on listener faces via time-shifted SyncNet contrast (details in Appendix A.2). Lower values indicate that listener mouth movements remain natural and independent of acoustic timing, rather than being falsely modulated by speech.

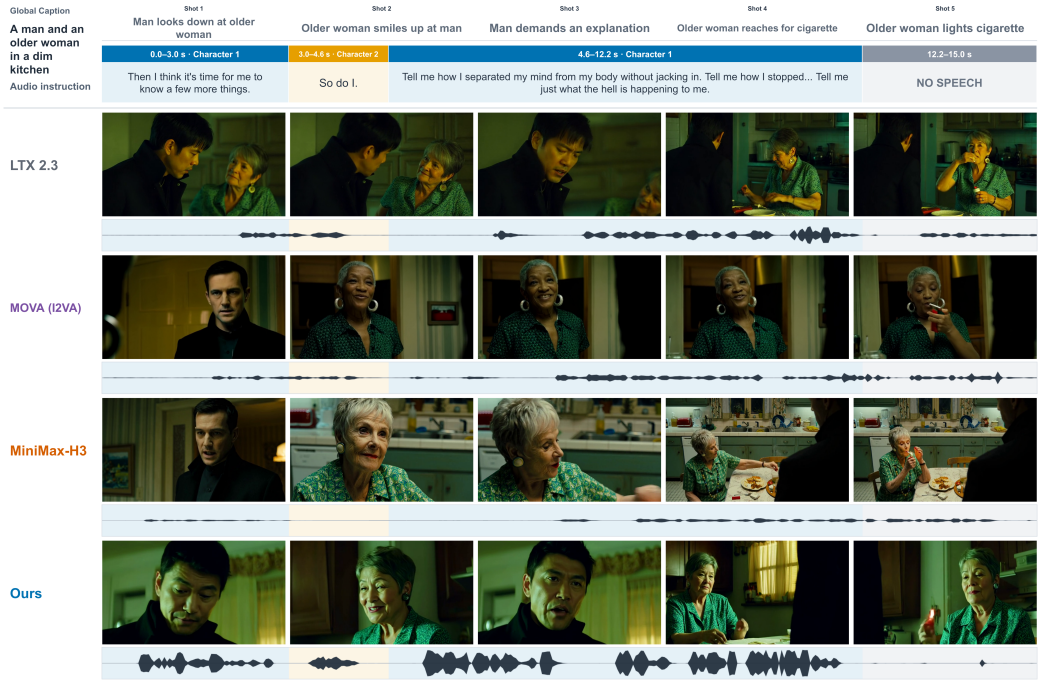


Figure 4: Qualitative comparison on a five-shot screenplay. Columns present shot prompts and the target dialogue schedule; rows show keyframes and synthesized audio waveforms across models. Blue/orange highlights mark each character's speech windows, and gray denotes silence. DÉCOUPAGE adheres to the multi-shot structure across transitions, respects the Shot-4 listener, and correctly synthesizes the Shot-5 slience and lighter flick, achieving superior audio-visual alignment.

### 4.3 QUANTITATIVE RESULTS

**Comparison on 15-Second Generations.** Table 1 shows that DÉCOUPAGE achieves the strongest screenplay control and character voice consistency among the evaluated methods on the 15-second benchmark. It achieves near-perfect adherence to the requested edits, with 0.99 Cut F1 and 0.05 Count MAE. It also obtains the lowest WER (4.59%), the highest containment (93.01%), and the lowest silence leakage (1.12%). Beyond temporal control, DÉCOUPAGE achieves the highest Role Consistency (0.53) and Role Margin (0.29), maintaining consistent character voices across conversations. These gains preserve competitive audiovisual quality. DÉCOUPAGE also improves AudioBox P-Q from 6.18 to 6.91 and Sync-C from 2.52 to 3.59. Although MiniMax-H3 achieves a higher Sync-C score (5.00), the supplementary video comparisons show that DÉCOUPAGE provides satisfactory lip synchronization for the cinematic video generation.

**Scaling to Minute-Long Narratives.** When scaling the generation horizon to 60 seconds, DÉCOUPAGE robustly maintains its high-fidelity temporal control and narrative coherence. As shown in Table 1, it preserves an impressive 0.76 Cut F1 and 91.12% containment with a mere 4.46% silence leakage. In stark contrast, standard baselines struggle severely over extended durations—for instance, LTX-2.3 Base degrades to 0.05 Cut F1 and 11.88% containment, while its silence leakage surges to 32.71%. Furthermore, DÉCOUPAGE retains strong character voice consistency and separation, achieving 0.57 Role Consistency and 0.22 Role Margin. These results confirm that our framework's screenplay adherence and acoustic grounding scale gracefully to continuous, minute-long cinematic scenes without drifting or collapsing.

### 4.4 QUALITATIVE RESULTS

Figure 4 compares DÉCOUPAGE with LTX-2.3 Base (HaCohen et al., 2026), MOVA (SII-OpenMOSS Team et al., 2026), and MiniMax-H3 (MiniMax, 2026) on the same five-shot screenplay. The example contains two alternating speakers, a long third dialogue turn that spans two visual shots, and a final 2.8-second no-speech interval. The baseline rows show weaker adherence to the requested shot sequence and poorer control of dialogue timing, consistent with the dataset-

| Configuration | Video Quality | | Audio Quality | | | | AV Alignment | | Efficiency |
|---|---|---|---|---|---|---|---|---|---|
| | Cut F1 ↑ | Imaging ↑ | WER ↓ | Cont. ↑ | Leak ↓ | Sim. ↑ | Bias (L) ↓ | Sync-C (S) ↑ | Time (s/step) ↓ |
| *Dual-schedule temporal CA* | | | | | | | | | |
| w/o Temporal CA | 0.34 | 0.65 | 6.57 | 39.34 | 25.98 | 0.39 | 2.31 | 3.16 | **12.81** |
| *Directorial Lip-Sync Mech.* | | | | | | | | | |
| w/o Mechanism | **0.99** | **0.66** | 4.38 | 90.89 | **0.74** | 0.55 | 1.78 | **3.53** | **12.81** |
| w/ Mechanism (**Ours**) | **0.99** | **0.66** | **4.16** | **92.83** | 0.77 | **0.56** | **0.33** | 3.42 | 13.01 |

Table 2: Ablation study on core components. Evaluated on a fixed subset of the 15-second conversational benchmark. Best values within each block are in **bold**. Notably, the lip-sync mechanism is entirely training-free: both configurations in the second block operate on the exact same checkpoint. Runtime denotes seconds per denoising step measured on a single NVIDIA H20 GPU.

level results in Table 1. In contrast, DÉCOUPAGE faithfully follows the screenplay's visual-shot and acoustic-event schedules: it reproduces all five requested compositions, carries the third utterance across the cut to the Shot-4 listener without driving the listener's lips, and preserves the Shot-5 no-speech interval while generating the lighter flick. Qualitative results for 60-second generations are presented in Appendix A.1, with more qualitative examples provided in the supplementary materials.

### 4.5 Ablation Study

Table 2 examines both components on a fixed subset of 136 of the 200 screenplays in the 15-second benchmark, where the lip-sync mechanism is applicable. All ablation configurations use the same subset, while Table 1 reports results on the full benchmark. Removing dual-schedule temporal cross-attention reduces Cut F1 from 0.99 to 0.34 and dialogue containment from 92.83% to 39.34%, demonstrating the importance of independently timed text conditioning for shot and dialogue scheduling. The directorial lip-sync mechanism addresses a complementary aspect of control: whether active speech should drive visible articulation. Compared with native audiovisual coupling, it reduces listener TemporalBias from 1.78 to 0.33, an 81.5% relative reduction, while Imaging Quality, speaker Sync-C, and audio metrics remain broadly stable. These results indicate reduced unwanted audiovisual coupling on listener shots while retaining visible-speaker synchronization. A matched qualitative comparison is provided in Appendix A.3, with videos in the supplementary materials. To assess the computational cost of these gains, we compare inference runtime with and without the directorial lip-sync mechanism using the same model checkpoint. On a single NVIDIA H20 GPU, enabling the mechanism increases the per-step runtime from 12.81 to 13.01 seconds (Table 2), corresponding to an additional 0.20 seconds (1.56%). This modest overhead requires no additional training or learned parameters.

## 5 Conclusion

DÉCOUPAGE enables independent handling of visual shot cuts and continuous cinematic audio through dual-schedule temporal cross-attention. It orchestrates shot transitions and dialogue schedules via decoupled time schedules, while the directorial lip-sync mechanism leverages auxiliary non-speech acoustic states to suppress phantom articulation on silent listeners. Together, these designs naturally unlock cinematic reaction shots and sound-bridge effects for long-form video generation. Evaluations over 15-second and 60-second multi-shot conversational scenes demonstrate clear improvements in cut adherence, dialogue containment, and recurring-character voice consistency, yielding high-quality narrative multi-shot audiovisual outputs.

**Limitations and Future Work.** Our current formulation targets turn-taking dialogue and does not explicitly support simultaneous, overlapping speech from multiple characters. This limits performance for crowded group scenes where several actors speak concurrently. One potential path forward is to extend our acoustic track to model multiple co-active speech events within the same time interval. Moving forward, an exciting frontier is exploring real-time, directorial-level joint audiovisual generation to enable responsive, on-the-fly cinematic creation. In parallel, we envision developing assisted screenplay authoring tools to simplify high-level creative control and lower end-user barriers.

## AI USE STATEMENT

Generative AI tools were used to assist with literature organization and the drafting and revision of manuscript text, including improving clarity, grammar, and readability. All AI-assisted content was reviewed and edited by the authors to ensure that it accurately reflected the research conducted. The authors verified the cited sources, checked technical descriptions and implementation claims against the code, and reviewed the experiments and results included in the paper. All substantive research decisions and interpretations of the findings remained the responsibility of the authors. The authors take full responsibility for the accuracy, originality, and integrity of the final manuscript.

## REPRODUCIBILITY STATEMENT

To support reproducibility, we will release the full codebase and accompanying documentation upon publication. The release will include environment specifications and dependencies, model configurations, hyperparameters, random seeds, data preparation and splitting procedures, and training schedules. We will also provide the evaluation scripts and metric implementations used to produce the reported results, together with step-by-step commands for running the experiments. Hardware configurations and execution-time information will be documented to help researchers assess the computational requirements. These materials will enable independent verification of our findings and facilitate further research building on our work.

## REFERENCES

Yuheng Chen, Teng Hu, Yuji Wang, Qingdong He, Zhucun Xue, Qianyu Zhou, Xiangtai Li, Lizhuang Ma, Jiangning Zhang, and Dacheng Tao. CineDance: Towards next-generation multi-shot long-form cinematic audio-video generation. *arXiv preprint arXiv:2606.09639*, 2026.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Google DeepMind. Veo 3: Technical report. Technical report, Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio-video generation through cross-task synergy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16085–16095, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hu_Harmony_Harmonizing_Audio-Video_Generation_through_Cross-Task_Synergy_CVPR_2026_paper.html.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-Your-Avatar: Multi-talking-character video generation with dynamic 3d-mask-based embedding router. *arXiv preprint arXiv:2506.19833*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, et al. VBench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. ShotAdapter: Text-to-multi-shot video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28405–28415, 2025. doi: 10.1109/CVPR52734.2025.02645.

Guillaume Klein. Faster-whisper: Faster whisper transcription with CTranslate2. GitHub repository, 2023. URL https://github.com/SYSTRAN/faster-whisper.

Kling Team et al. Kling-Omni technical report. *arXiv preprint arXiv:2512.16776*, 2025. doi: 10.48550/arXiv.2512.16776. URL https://arxiv.org/abs/2512.16776.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director. Official release announcement, February 5, 2026. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be.

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. LatentSync: Taming audio-conditioned latent diffusion models for lip sync with SyncNet supervision. *arXiv preprint arXiv:2412.09262*, 2024. doi: 10.48550/arXiv.2412.09262. URL https://arxiv.org/abs/2412.09262.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. VideoDirectorGPT: Consistent multi-scene video generation via LLM-guided planning. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2309.15091.

Haohe Liu, Gael Le Lan, Xinhao Mei, Zhaoheng Ni, Anurag Kumar, Varun Nagaraja, Wenwu Wang, Mark D. Plumbley, Yangyang Shi, and Vikas Chandra. SyncFlow: Toward temporally aligned joint audio-video generation from text. *arXiv preprint arXiv:2412.15220*, 2024. doi: 10.48550/arXiv.2412.15220. URL https://arxiv.org/abs/2412.15220.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. JavisDiT: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2503.23377.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. MAVIN: Multi-shot audio-visual generation with narrative control. *arXiv preprint arXiv:2606.29473*, 2026b.

Bingqi Ma, Linlong Lang, Ming Zhang, Dailan He, Xingtong Ge, Yi Zhang, Guanglu Song, and Yu Liu. Improving joint audio-video generation with cross-modal context learning. *arXiv preprint arXiv:2603.18600*, 2026.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. HoloCine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

Yihao Meng, Zichen Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yue Yu, Hanlin Wang, Haobo Li, Jiapeng Zhu, Yanhong Zeng, Xing Zhu, Yujun Shen, Qifeng Chen, and Huamin Qu. CausalCine: Real-time autoregressive generation for multi-shot video narratives. *arXiv preprint arXiv:2605.12496*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. MiniMax Research blog, July 2026. URL https://www.minimax.io/blog/minimax-h3.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 484–492, 2020. doi: 10.1145/3394171.3413532.

Chunyu Qiang, Jun Wang, Xiaopeng Wang, Kang Yin, Yuxin Guo, Xijuan Zeng, Nan Li, Zihan Li, Yuzhe Liang, Ziyu Zhang, Teng Ma, Yushen Chen, Zhongliang Liu, Feng Deng, Chen Zhang, and Pengfei Wan. MM-Sonate: Multimodal controllable audio-video generation with zero-shot voice cloning. *arXiv preprint arXiv:2601.01568*, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

Team Seedance et al. Seedance 1.5 Pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025. doi: 10.48550/arXiv.2512.13507. URL https://arxiv.org/abs/2512.13507.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. doi: 10.48550/arXiv.2604.14148. URL https://arxiv.org/abs/2604.14148.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. AudioBox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Xihua Wang, Ruihua Song, Chongxuan Li, Xin Cheng, Boyuan Li, Yihan Wu, Yuyue Wang, Hongteng Xu, and Yunfeng Wang. Animate and sound an image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23369–23378, 2025. doi: 10.1109/CVPR52734.2025.02176.

Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, and Yu Qiao. ShotDirector: Directorially controllable multi-shot video generation with cinematographic transitions. *arXiv preprint arXiv:2512.10286*, 2025. doi: 10.48550/arXiv.2512.10286. URL https://arxiv.org/abs/2512.10286.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Yang Zhao, Shanchuan Lin, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. VideoAuteur: Towards long narrative video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19163–19173, 2025. doi: 10.1109/ICCV51701.2025.01781.

Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. UniAVGen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. MuseTalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. *arXiv preprint arXiv:2410.10122*, 2024. doi: 10.48550/arXiv.2410.10122. URL https://arxiv.org/abs/2410.10122.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. MovieDreamer: Hierarchical generation for coherent long visual sequences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7d4647db960780c8b5a4e7d9e4a58d68-Abstract-Conference.html.

Yang-Hao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqin Zhou, Yudong Li, Jiajun Xu, Jingyun Liao, Yi-Ming Cheng, Xuefeng Chen, Xian-Ling Mao, and Yousheng Feng. MTAVG-Bench: A diagnostic benchmark for multi-talker dialogue-centric audio-video generation. *arXiv preprint arXiv:2602.00607*, 2026.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110315–110340, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c42f76b2ec3b3acff7c7b31f6486d0c9-Abstract-Conference.html.




Figure 5: Qualitative example of a 60-second generation. Eight representative frames show a two-character exchange unfolding across multiple shots in a trumpet shop. Beneath each row, the generated audio waveform is aligned with the corresponding shot sequence. The colored strips indicate the screenplay-specified speaking role (C1 or C2), while gray segments denote intervals with no requested speech. The example visualizes coordinated shot transitions, character turns, and speech–silence timing over a full-minute sequence.

# A Supplementary Results

## A.1 Minute-Long Qualitative Results

Figure 5 illustrates our model's ability to generate a 60-second multi-shot audiovisual sequence. More examples are provided in the supplementary materials.

## A.2 Metric Definitions and Evaluation Protocol

**Dialogue accuracy.** For video $i$, let $n_i$ be the number of reference words and let $s_i$, $d_i$, and $u_i$ denote substitutions, deletions, and insertions between the screenplay dialogue and the Faster-Whisper transcript. We report corpus-level word error rate as

$$\mathrm{WER} = 100\,\frac{\sum_i (s_i + d_i + u_i)}{\sum_i n_i}. \tag{8}$$

**Silence leakage.** Let $S_i$ be the union of requested silence intervals in video $i$, $S_i^{-\tau}$ the union after removing $\tau$ seconds from each end of each interval, and $W_i$ the union of recognized word intervals. With $\mu$ denoting duration, we define

$$\mathrm{SL}_\tau = 100\,\frac{\sum_i \mu\big(S_i^{-\tau} \cap W_i\big)}{\sum_i \mu(S_i)}, \qquad \tau = 0.2\,\mathrm{s}. \tag{9}$$

Intervals shorter than $2\tau$ have no remaining interior, while the denominator retains the original requested silence duration. Using word intervals avoids counting pauses enclosed within longer recognition segments.

**Text-aligned containment.** For turn $t$, let $[a_t, b_t]$ be the allowed speech window and $[\hat{a}_t, \hat{b}_t]$ its boundaries obtained from transcript alignment. The total boundary overflow is

$$o_t = \max(0, a_t - \hat{a}_t) + \max(0, \hat{b}_t - b_t). \tag{10}$$

For the set $T$ of all target turns and its validly aligned subset $A$, containment at tolerance $\delta$ is

$$\mathrm{Containment}_\delta = \frac{100}{|T|} \sum_{t \in A} \mathbf{1}[o_t \le \delta]. \tag{11}$$

Missing or invalid alignments remain in the denominator and count as failures; unused time within the allowed window incurs no penalty. We use $\delta = 0.2\,\mathrm{s}$ for the 15-second evaluation dataset and $\delta = 0.4\,\mathrm{s}$ for the 60-second evaluation dataset.

**Recurring-role voice consistency.** Let $e_{ict}$ be the frozen ECAPA-TDNN embedding of an eligible turn $t$ for character $c$ in video $i$. Let $P_{ic}$ contain all unordered pairs of distinct eligible turns for that character, $C_i$ the characters with at least one such pair, and $\mathcal{V}$ the videos with at least one eligible character. We compute

$$\text{Consistency} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\frac{1}{|C_i|}\sum_{c\in C_i}\frac{1}{|P_{ic}|}\sum_{(t,u)\in P_{ic}}\frac{e_{ict}^{\top}e_{icu}}{\|e_{ict}\|_2\|e_{icu}\|_2}. \tag{12}$$

Timestamp-window crops use the requested speech windows and their screenplay character labels. Role Margin is computed within each eligible video as the difference between same-role and different-role similarity, rather than by subtracting independently supported aggregate means.

**Audio-Video Alignment** For Audio-Video alignment, prompt speech timestamps are intersected with requested shot intervals and visible roles to form listener and visible-speaker windows of at least 2.2 seconds. A single S3FD track must cover at least 95% of each continuous window. Imaging Quality is computed on complete 15-second clips. Shot transitions are detected at a fixed threshold of 0.18 and matched one-to-one to requested cuts within 0.5 seconds.

**TemporalBias formulation.** To measure soundtrack-specific temporal coupling on listener/off-screen windows, we define an Audio-Lock contrast using SyncNet confidence. For each window $m$, let $C_m(v_m, a_m)$ denote the near-lag SyncNet confidence between the detected facial track $v_m$ and its aligned audio $a_m$. We construct temporally displaced controls by circularly shifting $a_m$ by a fraction $\delta$ of the window duration, where $\mathcal{D} = \{0.25, 0.375, 0.5, 0.625\}$. The contrast is

$$\Delta C_m = C_m(v_m, a_m) - \operatorname*{median}_{\delta\in\mathcal{D}} C_m(v_m, \mathrm{shift}(a_m, \delta))\,. \tag{13}$$

We then aggregate the contrast magnitude over the $N$ listener/off-screen windows:

$$\text{TemporalBias} = \frac{1}{N}\sum_{m=1}^{N}|\Delta C_m|. \tag{14}$$

If a listener's facial dynamics exhibit no soundtrack-specific temporal preference, $\Delta C_m$ should be close to zero. Accordingly, a lower TemporalBias indicates weaker unwanted audiovisual temporal coupling and is better.

### A.3 Matched Qualitative Directional Lip-Sync Mechanism Case

Figure 6 compares matched trajectories without and with the directional Lip-sync mechanism at three aligned timestamps near the end of Shot 2. The prompt, seed, checkpoint, sampling settings, and native V2A path are fixed; only the inference-time mechanism setting differs. Character 2 remains visible while Character 1 speaks off screen from 5.80 seconds until the cut at 6.50 seconds. All three displayed timestamps lie inside this interval. In these frames, the output without the mechanism moves from an open laugh toward a narrower mouth configuration during the off-screen line, whereas the matched output with the mechanism remains in an open-laugh configuration. Videos can be found in the supplementary materials.

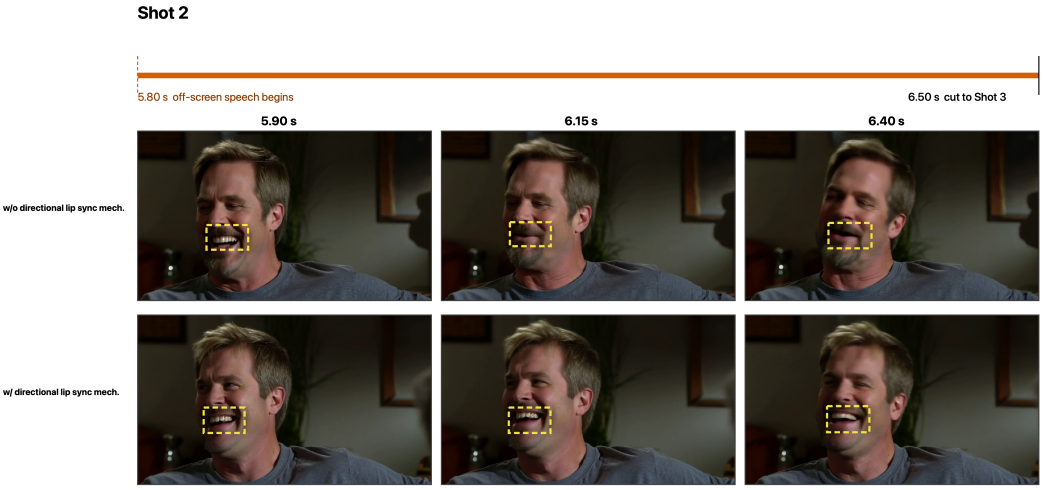


Figure 6: Matched comparison of the Directorial Lip-Sync Mechanism near the end of Shot 2. The first row shows results without the mechanism, and the second row shows results with it, at 5.90, 6.15, and 6.40 seconds. Yellow dashed boxes highlight the mouth region.