RAAVGEN: ROLLOUT-AWARE LONG-FORM AUDIO-VIDEO NARRATIVE GENERATION FROM GLOBAL SCRIPTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating long-form audio-video narratives from a complete script requires tracking what has already been generated and continuing unfinished content across generation boundaries. Segment-specific prompting workflows assign this responsibility to separately prepared local prompts. We introduce a rollout-aware global script framework that jointly generates audio and video segment by segment while retaining the complete script throughout generation. Our key insight is that visual continuity and script localization require different temporal contexts. We therefore pair a short video history that captures the recent visual state with a longer audio history that helps the model locate preceding speech in the script and determine what to generate next. This asymmetric design preserves extended spoken context while limiting the token cost of dense video history. For multi-shot narratives, sparse keyframes from completed shots provide earlier scene and character appearance, while relative shot timestamps guide shot transitions. Together, these conditions allow generation boundaries to fall within shots and utterances. To address limited long-form training data, we combine complementary supervision with staged training to integrate face and voice conditioning, shot transitions, and long-sequence continuation. We further construct pseudo-long scripts by extending single-shot scripts with speech text and multi-shot scripts with additional shot descriptions, training script localization while retaining the original audio-video supervision. Experiments demonstrate long-form multi-shot generation from a complete script without separately prepared local prompts, alongside support for optional face and voice personalization.

1 INTRODUCTION

Generating audio-video narratives from textual scripts has broad applications in filmmaking and digital storytelling Meng et al. (2026); Zhou et al. (2024), as well as virtual character creation Lin et al. (2025). Recent advances in joint generation have enabled models to synthesize realistic visual content together with temporally aligned speech and sound from textual descriptions Low et al. (2025); Liu et al. (2026a); HaCohen et al. (2026). These advances motivate the generation of longer narratives spanning multiple shots, where characters must retain their visual identities and voices as scenes and dialogue evolve. Such narratives also require events to follow the script, with speech continuing naturally across successive segments. The central challenge is to satisfy these requirements jointly over extended durations, maintaining both the consistency of individual characters and the coherence of the overall narrative.

Existing approaches to narrative generation broadly follow two paradigms, jointly generating the entire sequence within a single temporal window or extending it autoregressively through successive clips. Within the first paradigm, general joint audio-video generators primarily focus on perceptual quality and audiovisual synchronization, with limited explicit support for multi-shot narrative control Low et al. (2025); Zhang et al. (2026a); Wang et al. (2025). More recent methods introduce dedicated mechanisms for multi-shot narrative control and temporal script alignment Liu et al. (2026c;d). However, generating longer sequences within a single temporal window increases computational and memory demands Meng et al. (2026), motivating generation through successive segments. Autoregressive audio-video generation offers an alternative by extending sequences through

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.

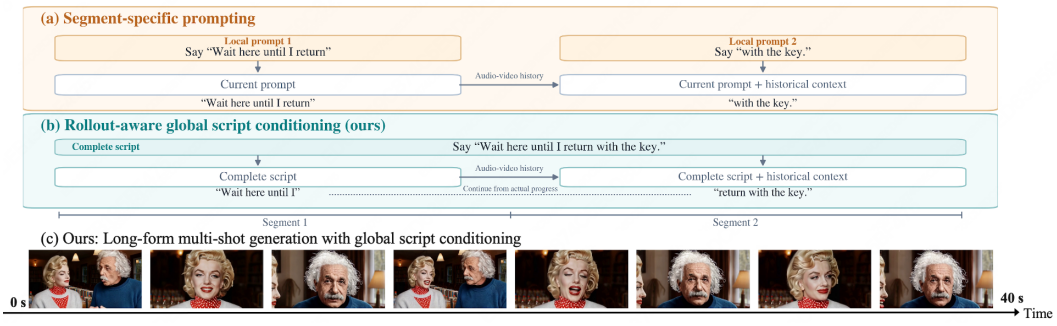

Figure 1: Comparison of prompting strategies and long-form generation. (a) Segment-specific prompting assigns a separate local prompt to each generation segment, requiring the dialogue to follow a predefined text partition. (b) Our method retains the complete script across segments and incorporates prior audio-video context, allowing unfinished dialogue to continue according to the actual generation progress. (c) Conditioned on a global script, our method generates a coherent 40-second multi-shot audio-video sequence, maintaining character identity and voice timbre across distant shots while preserving background consistency when scenes reappear.

successive segments conditioned on historical context or memory Zhang et al. (2026b); Huang et al. (2026); Li et al. (2026a). In representative segment-based or shot-based workflows, local prompts specify the narrative content to be generated at each step. As illustrated in Fig. 1(a), continuing dialogue across two segments then requires separate prompts specifying the content of each segment. Following a complete script across segments requires the generator itself to determine what has already been realized and how unfinished speech or actions should continue.

To address this challenge, we introduce a rollout-aware global script framework for long-form audio-video narrative generation. Given a complete script and optional face and voice references, our model generates the narrative segment by segment while keeping the full script available throughout generation, as illustrated in Fig. 1(b). When a segment ends within an utterance, the next step must locate the remaining content and continue from the state already reached. We use asymmetric audio-video history to support this process in both single-shot and multi-shot generation. A short video history provides the visual state near the segment boundary, while a longer audio history preserves preceding speech as evidence for script localization and dialogue continuation. In our latent representation, audio history uses fewer tokens per unit of time than video history, allowing us to retain a longer record of spoken content while limiting the cost of dense visual context. For multi-shot narratives, we further retain keyframes from completed shots as sparse visual memory. These keyframes complement the short video history with earlier scene and character appearance information, supporting consistency when characters or settings reappear. We also express shot timestamps relative to the start of the current generation segment. The relative shot timestamps guide shot changes, while the histories and visual memory provide context for continuing the narrative. Together, these conditions support script following and continuation across segments whose boundaries need not align with shots or utterances.

Training a model for this setting requires supervision for face and voice conditioning, script following, and continuation across shots. These signals are difficult to obtain together, since long multi-shot recordings are scarce and changes in viewpoint and speaker complicate the extraction of reliable face and voice references. We therefore develop a compositional long-horizon training framework that learns from complementary data sources. Two-speaker single-shot conversations support face and voice conditioning, while long single-shot sequences and multi-shot videos provide supervision for temporal continuation and shot transitions. We further distinguish between learning to find the relevant content in a long script and learning to continue generation over many steps. For script localization, we construct pseudo-long scripts by adding speech text before and after single-shot scripts and additional shot descriptions before and after multi-shot scripts. The original audio-video supervision remains unchanged, allowing the model to learn script localization and continuation within longer scripts without requiring matching full-length recordings. We organize training in three stages, first learning face and voice conditioning from short clips, then learning shot transitions from short multi-shot sequences with timestamped scripts, and finally learning long-

sequence continuation from a large amount of long single-shot data and a small amount of long multi-shot data. Together, these strategies support script localization within longer textual contexts and sustained continuation over more generation steps.

From a complete script and optional face and voice references, our framework generates coherent 40-second multi-shot audio-video narratives with multiple speakers. This enables narrative content to be specified independently of generation boundaries, reducing the need to manually divide and coordinate dialogue and actions across segments.

Our contributions are summarized as follows.

- We study long-form personalized audio-video narrative generation from a complete script, with shot timestamps for multi-shot narratives and generation boundaries that need not align with shots or utterances.
- We introduce rollout-aware global script conditioning with asymmetric audio-video history for boundary continuity and script localization, complemented by sparse visual memory for cross-shot consistency and relative shot timestamps for shot scheduling.
- We develop a compositional training framework that combines complementary supervision, pseudo-long script augmentation, and staged training to integrate personalization, long-sequence continuation, and multi-shot generation.

2 Related Work

2.1 Joint Audio-Video Generation

Joint audio-video generation synthesizes visual and acoustic content within a shared generative process. MM-Diffusion Ruan et al. (2023) explores coupled diffusion, while Seeing and Hearing connects pretrained generators through latent alignment Xing et al. (2024). MM-LDM Sun et al. (2024) and AV-DiT Wang et al. (2024) further investigate multimodal latent representations and transformer architectures. Recent methods, including JavisDiT Liu et al. (2026a), JavisDiT++ Liu et al. (2026b), Ovi Low et al. (2025), and LTX-2 HaCohen et al. (2026), improve generation quality and synchronization through cross-modal modeling. Ovi uses twin diffusion backbones with cross-modal fusion, whereas JavisDiT++ combines joint attention with modality-specific experts and temporally aligned positional encoding.

2.2 Long-Form and Multi-Shot Narrative Generation

Narrative generation requires temporal continuity and consistency across shots. VideoDirectorGPT Lin et al. (2023), VideoStudio Long et al. (2024), and MovieDreamer Zhao et al. (2025) organize stories through scene plans or hierarchical representations, while StoryDiffusion Zhou et al. (2024) emphasizes visual consistency. HoloCine Meng et al. (2026) jointly synthesizes multiple shots through window cross-attention and sparse inter-shot attention.

Alternatively, successive segments can be generated using historical context. StreamingT2V Henschel et al. (2025) propagates context for long video generation, while StoryMem Zhang et al. (2025), OneStory An et al. (2026), and Memento Wei et al. (2026) investigate cross-shot memory. Joint audio-video methods further address appearance, voice, and audiovisual continuity. UnityShots Huang et al. (2026) combines long-term and short-term visual memory with a speaker anchor, while Vorch-Director Zhang et al. (2026b) uses noise-aware history corruption and a clean reference sink to mitigate accumulated errors. JoyAI-Echo Li et al. (2026a) combines audiovisual memory with training for extended generation. In these multi-shot audio-video workflows, local prompts specify the content of the current segment or shot. We instead study continuation from a complete script, reducing the need to construct separate prompts for successive segments.

2.3 Temporal Control and Script Progress Tracking

Temporal control aligns textual instructions with their intended intervals. MinT Wu et al. (2025) uses time-aware positional encoding, Presto Yan et al. (2025) adopts segmented cross-attention, and

ShotAdapter Kara et al. (2025) introduces shot-aware conditioning. For joint audio-video generation, MAVIN Liu et al. (2026c) controls shot transitions and dialogue intervals through boundary-aware attention. TCR Liu et al. (2026d) aligns script timing with both modalities and allows dialogue to remain active across shot transitions. These methods primarily address temporal alignment within a generation pass.

Progress tracking is also studied across successive generation chunks. StreamChar Tian et al. (2026) uses a jointly trained LLM orchestrator to derive continuous audio conditions from text and historical audio, with a progress-aware pointer supporting transcript alignment during rollout training. Vorch-Streamer Han et al. (2026) supplies each causal block with temporally aligned planning features from an external speech language model. Both methods primarily target continuous character or avatar generation, focusing on speech progression and long-term consistency. Our setting additionally requires following a multi-shot narrative script that coordinates shot transitions, dialogue, and visual events. We combine relative temporal anchoring with prior audio-video context to track script progress and continue unfinished events across segments without a speech-planning module.

2.4 Personalization and Training for Long Narratives

Personalized audio-video generation controls both visual identity and voice. OmniCustom Li et al. (2026c) studies joint identity and timbre customization, while DreamID-Omni Guo et al. (2026) addresses identity and voice binding through a unified framework with progressive multi-task training. Identity as Presence Chen et al. (2026a) investigates appearance and voice personalization, and MMControl Li et al. (2026b) broadens multimodal control.

MAVINSet Liu et al. (2026c) and CineDance-1M Chen et al. (2026b) provide structured annotations for multi-shot audio-video generation. Our framework combines complementary supervision from conversations, long single-shot sequences, and multi-shot videos with pseudo-long script augmentation and three-stage training. The augmentation expands the textual context while retaining the original audio-video supervision, supporting script localization within longer narratives.

3 Method

Given a complete script C and optional face and voice references $\mathcal{R}$, our model generates successive audio-video segments $Y_i = (V_i, A_i)$. At step i, generation is conditioned on C_i, asymmetric audio-video history $\mathcal{H}_i$, and sparse visual memory $\mathcal{S}_i$. Figure 2 shows how history and memory are extracted from the generated sequence and combined with the script and optional references to produce the next segment. For single-shot generation, $C_i = C$ and $\mathcal{S}_i$ is empty. For multi-shot generation, C_i retains the full script with shot timestamps relative to the segment start τ_i. The requested duration D_i determines target latent lengths. Generation is autoregressive across segments, while audio and video within each segment are jointly generated through rectified flow. Segment boundaries need not align with shot transitions or the ends of utterances.

3.1 Multimodal Condition Encoding

Frozen video and audio VAEs encode targets, references, and histories into their respective latent spaces. A frozen Qwen3-VL Bai et al. (2025) encoder processes the complete script and available face references. The image placeholders `@CTX_IMG_k` are grouped into a single prefix before the entire script and replaced by the corresponding face images. Face references also pass through the video VAE Wan et al. (2025) and enter the joint sequence as visual context tokens. Voice references are encoded by the audio VAE and serve as timbre conditions.

The script describes characters, speakers, and their dialogue in natural language. For multiple speakers, paired face and voice references follow the same order, with the first speaker's conditions placed first. This order remains fixed across generation steps, while the model learns reference-to-character correspondence from the script and audio-video context. For multi-shot generation, sparse visual memory enters only through the video VAE as sequence context tokens, without text-encoder image placeholders. Encoder configurations and preprocessing are provided in the supplementary material.

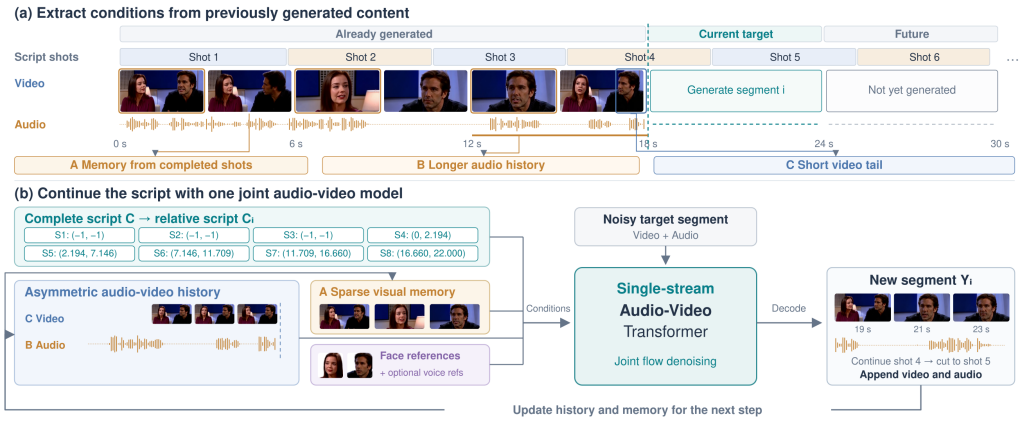

Figure 2: Overview of our framework. (a) At the current generation boundary, sparse visual memory (A), longer audio history (B), and a short video tail (C) are extracted from previously generated content. Both histories end at this boundary, marked by the vertical dashed line at 18 s. (b) A single-stream audio-video transformer jointly generates the next segment conditioned on the complete script, reference conditions, and the extracted history and memory. All shot descriptions are retained, with their timestamps updated relative to the current generation boundary. Only timestamps are shown for brevity. Newly generated content updates history and memory for subsequent steps. Single-shot generation omits shot timestamps and sparse memory.

3.2 Rollout-Aware Global Script Conditioning

The complete script remains available at every generation step, even when its content extends beyond the current target window. The model uses history to locate unfinished content and continue it without separately prepared local prompts.

Asymmetric Audio-Video History. As shown in Fig. 2(a), we use $\mathcal{H}_i = (H_i^v, H_i^a)$, pairing a short video tail with a longer audio history that ends at the same generation boundary. The video tail provides the visual state near the segment boundary, while preceding speech helps locate progress within the complete script and continue unfinished dialogue.

For latent temporal lengths T_v and T_a and a video spatial grid of size $H_v \times W_v$, video token counts scale with $T_v H_v W_v$, while audio token counts scale with T_a. A longer audio history therefore preserves more spoken context at a lower token cost than an equally long dense video history in our representation. Both single-shot and multi-shot continuation use this asymmetric history.

Sparse Visual Memory. For multi-shot narratives, we randomly select one frame from each completed shot and encode the selected frames as sparse visual memory $\mathcal{S}_i$. The memory is ordered chronologically and updated as shots are completed. It complements the short video history with earlier scene and character appearance, providing context when a character or setting reappears. Target tokens attend to these memory frames alongside recent history.

Relative Shot Timestamps. For multi-shot narratives, we represent the complete script as an ordered sequence $C = \big((s_j, e_j, c_j)\big)_{j=1}^{J}$, where c_j describes shot j over the interval $[s_j, e_j)$. At target start τ_i, we convert each pair of timestamps into local coordinates,

$$\mathcal{T}_{\tau_i}(s_j, e_j) = \begin{cases} (-1, -1), & e_j \le \tau_i, \\ (0, e_j - \tau_i), & s_j < \tau_i < e_j, \\ (s_j - \tau_i, e_j - \tau_i), & s_j \ge \tau_i. \end{cases} \tag{1}$$

The output is an ordered timestamp pair, with $(-1, -1)$ marking a completed shot. For a shot that spans the generation boundary, the local start timestamp is set to 0, while its end timestamp indicates the remaining duration. The updated script C_i retains all shot descriptions and their relative timestamps without truncating them to the current generation window. Figure 2(b) illustrates this update for completed, ongoing, and future shots at a continuation step. Relative timestamps guide the shot

schedule, while the script and audio-video history provide evidence of spoken-content progress. Single-shot scripts contain no timestamps, with audio-video history providing cues for script localization and continuation.

3.3 Single-Stream Audio-Video Flow Transformer

Joint token processing. Modality-specific projections and refiners map audio, video, and script features into a shared single-stream DiT with separate audio and video output heads. Reference, memory, and history tokens attend only within the condition block, preventing them from attending to noisy targets. Target and script tokens attend to all valid tokens, without a token-level causal mask or shot-specific restrictions on script access. The token layout and attention mask are detailed in the supplementary material.

Reference slots and positional encoding. We apply rotary positional encoding over three coordinate axes. Video tokens use temporal and spatial coordinates (p_t, p_h, p_w), while script and audio tokens use (p, p, p), repeating their sequence or temporal position across all three axes.

In the DiT positional coordinate system, reference conditions are assigned coordinates after the script range. This coordinate assignment is independent of the image-prefix arrangement in the text encoder. Let p_{end} denote the last script position, K the number of reference slots, and M the number of sparse visual memory frames. Paired face and voice conditions share a slot, while a face-only or voice-only condition occupies its own slot. We assign context positions as

$$b_j = p_{\mathrm{end}} + j\Delta_{\mathrm{ctx}}, \qquad j = 1, \ldots, K + M. \tag{2}$$

For reference pair k, face tokens use coordinates (b_k, h, w), while all tokens of the paired voice reference use (b_k, b_k, b_k). The shared temporal position associates face and voice conditions while preserving separate token groups. Sparse visual memory occupies subsequent visual positions without corresponding audio reference groups.

The video timeline, comprising history and target tokens, starts after the context positions with an additional gap Δ_{seq}. The audio timeline is shifted to align its target start with the video target, as described below. Both Δ_{ctx} and Δ_{seq} provide positional separation between token groups without inserting additional tokens.

Aligning unequal audio-video histories. The asymmetric histories in Sec. 3.2 have different durations and latent sampling rates. We therefore align the start of the two targets in a shared latent-time coordinate system. For video frame rate f, we use $\alpha_v = f_0/f$ with a reference frame rate $f_0 = 15$ FPS. This reference rate sets the coordinate scale, not the input frame rate. The audio step α_a is calibrated to the video latent-time scale, with the calibration procedure provided in the supplementary material. Let b be the video timeline origin after the reference and memory slots. The temporal coordinates are

$$\begin{aligned} p_v(j) &= b + j\alpha_v, \\ p_a(\ell) &= b + n_v\alpha_v + (\ell - n_a)\alpha_a. \end{aligned} \tag{3}$$

Here n_v and n_a denote the numbers of video and audio history latents, respectively. With zero-based indexing, the first target tokens occur at indices n_v and n_a in their respective sequences. Their temporal coordinates coincide at $p_v(n_v) = p_a(n_a) = b + n_v\alpha_v$, aligning the target starts despite different history lengths.

3.4 Compositional Training

Complementary supervision. Long multi-shot conversations are scarce, and changes in speaker and viewpoint complicate the extraction of reliable face and voice conditions. We therefore combine short single-person and two-person clips, long single-shot sequences, and multi-shot videos. These sources provide complementary supervision for personalization, continuous generation, and shot transitions, allowing the shared model to combine these capabilities.

Pseudo-long script augmentation. Long recordings with aligned scripts are limited. For continuation examples with audio history, we extend the original script with additional speech text for single-shot data and additional shot descriptions for multi-shot data. In both cases, we form $\widetilde{C} = C^- \oplus C \oplus C^+$, where C^- and C^+ contain the preceding and following script content, respectively, and $\oplus$ denotes concatenation. The original audio-video history and target remain unchanged.



Figure 3: Qualitative comparison of long-sequence multi-shot generation. The left and right groups show two cases at selected timestamps shared by all methods. Audio waveforms are colored according to the WavLM timbre similarity between each utterance and the first utterance of the same character (green: higher; red: lower). Our method better preserves character identity, speaker timbre, and scene consistency over time.

The model learns to locate the content corresponding to its history within the extended script and continue from that point, without requiring a matching full-length recording.

The three-stage training schedule, history-target sampling procedure, and joint flow matching objective are detailed in the supplementary material.

4 Experiments

We evaluate RaAVGen on long-form multi-shot audio-video generation and face and voice personalization. For comparison with existing methods, we use GPT to construct a test set of 50 text prompts, each describing an approximately 40-second audio-video narrative with 6-14 shots. These prompts serve as inputs to the evaluated generation models. Metrics are computed from the generated audio-video outputs using the evaluation procedures described in the supplementary material.

Compared methods. We group the compared methods into three categories according to their generation capability: (i) single-shot short-sequence generators, including Ovi Low et al. (2025) and DreamID-Omni Guo et al. (2026); (ii) multi-shot short-sequence generators, including TCR Liu et al. (2026d) and MiniMax-H3[1]; and (iii) the multi-shot long-sequence generator JoyAI-Echo Li et al. (2026a).

Since generating an entire 40-second sequence within a single generation segment is challenging for the evaluated methods, we adopt method-specific segmentation strategies while preserving the same script content and reference conditions. For TCR, MiniMax-H3, and JoyAI-Echo, we first decompose the complete timestamped script into shot-level prompts according to the annotated shot boundaries. We then greedily pack consecutive shots into generation segments under a common 10-second duration limit adopted for this evaluation. The timestamps are re-indexed relative to the beginning of each segment, and the number of generated frames is determined by the total annotated duration of the packed shots. The resulting segments are generated sequentially in temporal order and concatenated without temporal trimming. JoyAI-Echo additionally takes the output of the preceding segment as history when generating the next segment. Because TCR and Ovi do not support direct identity-reference injection, we condition TCR on the first frame of the first shot in each packed segment and Ovi on the first frame of each shot, both extracted from our generated sequence. Ovi generates each shot independently, trims it to the annotated duration, and concatenates the results chronologically. DreamID-Omni is likewise evaluated in its single-shot inference set-

[1] https://github.com/MiniMax-AI/MiniMax-H3

Table 1: Quantitative comparison with other methods for multi-shot long-sequence audio-visual generation. SB-F1 and WER are reported in percent, and SB-MAE is measured in seconds. DreamID and OVI are evaluated shot by shot and are excluded from the shot-boundary evaluation. OVI and TCR do not support reference-timbre conditioning and are therefore excluded from T-Sim. "–" denotes not applicable. The best and second-best results are shown in bold and underlined, respectively.

Method	**Cross-shot**		**Shot boundary**		**Inter-shot**		**TAMS**	**Audiobox**		**VBench**		**SyncNet**		**T-Sim**	**ID-Sim**	**WER**
	X-ID↑	X-T↑	F1↑	MAE↓	CISC↑	BISC↑	↑	CU↑	PQ↑	AES↑	IQ↑	C↑	D↓	↑	↑	↓
DreamID	0.392	0.784	–	–	0.626	0.642	0.928	5.808	5.793	0.522	0.644	3.885	10.343	0.764	0.321	5.42
OVI	0.373	0.875	–	–	0.713	0.591	0.917	6.397	6.872	0.484	0.631	4.227	10.926	–	0.395	8.52
TCR	0.325	0.887	52.60	0.174	0.702	0.663	**0.973**	7.053	7.735	**0.549**	0.657	5.292	9.565	–	0.280	2.71
H3	0.831	0.951	69.84	0.212	0.836	**0.692**	0.918	**7.398**	7.272	0.524	0.632	**5.904**	9.993	0.891	**0.697**	2.69
JoyAI-Echo	0.593	0.925	40.00	0.290	0.767	0.677	0.892	7.160	**7.738**	0.514	**0.680**	5.763	**9.265**	0.768	0.372	4.02
Ours (2s audio his.)	0.858	0.936	73.90	0.166	0.896	0.667	0.933	6.905	7.126	0.494	0.626	5.402	9.782	0.872	0.584	4.08
Ours (w/o memo.)	0.776	0.938	72.80	0.162	0.758	0.612	0.942	6.929	7.141	0.495	0.627	5.577	9.673	0.883	0.501	2.74
Ours	**0.874**	**0.956**	**74.60**	**0.159**	**0.924**	0.674	0.945	6.948	7.160	0.497	0.629	5.845	9.462	**0.894**	0.623	**2.64**

ting, with each annotated shot generated independently according to its duration and concatenated chronologically. In contrast, our method directly takes the complete timestamped script as input and performs segment-wise generation using the asymmetric audio-video history and sparse visual memory introduced in Sec. 3, with each segment lasting approximately eight seconds.

4.1 Comparison with Existing Methods

Quantitative comparison. As shown in Tab. 1, RaAVGen achieves strong cross-shot identity and timbre consistency, together with accurate shot-boundary placement, in long-form multi-shot audio-visual generation. In particular, its leading X-ID, X-T, and CISC results demonstrate stronger preservation of character identity, speaker timbre, and character appearance across distant shots. RaAVGen also follows the prescribed shot structure more accurately, achieving more reliable boundary detection and more precise boundary timing. Its strong T-Sim., ID-Sim., and WER results further indicate that long-range consistency is maintained without sacrificing reference fidelity or speech accuracy.

Several competing methods perform well on individual dimensions. For example, H3 achieves strong average background consistency and reference identity fidelity, while TCR and JoyAI-Echo perform favorably on selected perceptual-quality or synchronization metrics. Nevertheless, these methods do not provide the same balanced performance across cross-shot consistency, shot-structure accuracy, personalization, and speech generation. Overall, the results demonstrate that RaAVGen is more effective at jointly maintaining visual, acoustic, and structural coherence throughout long-form multi-shot generation.

Qualitative comparison. Fig. 3 presents frames sampled at timestamps shared by all methods from two long-sequence examples. Different baselines exhibit different forms of temporal inconsistency. Some methods alter facial attributes when a character reappears after several shots, whereas others show noticeable speaker-timbre drift across distant utterances. H3 preserves character identity relatively well, but the first example reveals localized changes in background composition and scene context, including noticeable variations in the visible room layout across shots.

In contrast, RaAVGen maintains more stable character appearances and speaker identities while preserving a coherent scene across changes in speaker, framing, and shot scale. These observations are consistent with its strong X-ID, X-T, CISC, ID-Sim., and T-Sim. results in Tab. 1. The waveform below each sequence shows speech activity, while its color encodes the WavLM timbre similarity between each utterance and the first utterance of the same character, with greener colors indicating higher similarity.

4.2 Ablation Studies

Audio history length. We compare our default configuration using 6 seconds of audio history with a variant using 2 seconds, while fixing video history at 4 latent frames and keeping all other settings

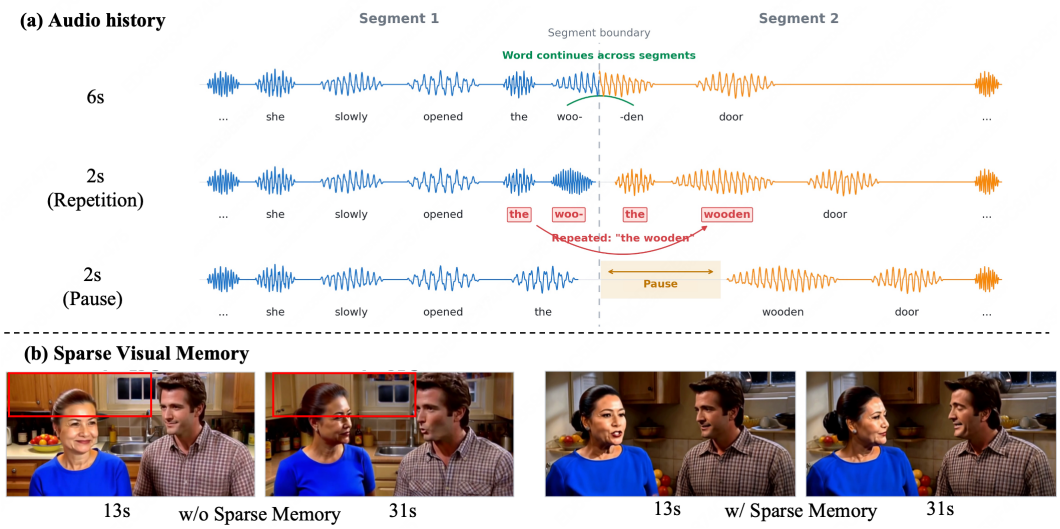

Figure 4: Qualitative ablations of audio history length and sparse visual memory. (a) With 6 seconds of audio history, an unfinished word continues smoothly across the segment boundary. With 2 seconds, continuation may exhibit repeated words or a pause before resuming speech. The waveforms are schematic. (b) Sparse visual memory preserves the background and scene layout across temporally distant shots, while removing it leads to background and camera-composition drift.

unchanged. As shown in Tab. 1, increasing audio history from 2 to 6 seconds reduces WER from 4.08% to 2.64%, supporting the use of longer speech context for script following.

Fig. 4(a) qualitatively compares speech continuation across segment boundaries under these two settings. Longer audio history provides more preceding speech context, helping the model locate its current position in the script and continue unfinished utterances across generation boundaries. Further analysis of audio history length is provided in Sec. E.2 of the supplementary material.

Sparse visual memory. We remove the sparse keyframes while keeping the remaining conditions unchanged. As reported in Tab. 1, removing sparse visual memory substantially degrades cross-shot identity consistency, inter-shot character and background consistency, and reference identity fidelity. In comparison, speaker consistency, speech accuracy, synchronization, and local perceptual quality remain relatively stable. This indicates that recent video history alone is insufficient to preserve character and scene information over long temporal intervals. This conclusion is further supported by Fig. 4(b). Without sparse visual memory, the background layout and camera composition change noticeably between the frames at approximately 13 and 31 seconds, whereas the full model retains a more consistent environment.

Additional quantitative ablations, implementation details, and qualitative comparisons are provided in the supplementary material.

5 Conclusion

We presented RaAVGen for generating extended audio-video narratives from a single complete script. Our approach uses generated history not only to preserve continuity but also to determine where generation should resume within the narrative at each successive generation step. Short visual context, longer speech context, and sparse memories of earlier shots serve distinct roles, allowing the model to carry forward the relevant information without retaining dense video history. Complementary training sources bring personalization, shot transitions, and sustained continuation into a shared model despite the limited availability of long-form supervision. Experiments demonstrate long-form multi-shot audio-video generation and face and voice personalization. This work offers a practical way to connect script-level authoring with segment-level generation, reducing the need to plan content around generation boundaries.

AI use statement We used generative AI tools to polish the manuscript and improve its clarity and readability. We also used GPT to generate text prompts for the test set. The authors reviewed the AI-assisted manuscript edits and take responsibility for the final manuscript and test prompts.

Ethics Statement Our work focuses on generating audio-video narratives purely from a technical standpoint, with no intent for malicious use. Nevertheless, we acknowledge the potential for misuse, such as the creation of fake videos. To mitigate this, it is imperative that synthetic videos are clearly labeled to reflect their artificial origin.

Reproducibility statement We provide a detailed description of the model architecture, training and inference procedures, and dataset construction methods in both the main paper and supplementary materials. The input caption is displayed in the demo.

REFERENCES

Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. In *CVPR*, pp. 16173–16184, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Yingjie Chen, Shilun Lin, Cai Xing, Binxin Yang, Long Zhou, Qixin Yan, Wenjing Wang, Dingming Liu, Hao Liu, Chen Li, et al. Identity as presence: Towards appearance and voice personalized joint audio-video generation. *arXiv preprint arXiv:2603.17889*, 2026a.

Yuheng Chen, Teng Hu, Yuji Wang, Qingdong He, Zhucun Xue, Qianyu Zhou, Jason 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*, 2026b.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, pp. 251–263. Springer, 2016.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

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

Menglin Han, Yang Ding, Yulei Lu, Haoran Yu, Xin Ma, Junyi Chen, Zhangkai Ni, Lin Ma, and Yaohui Wang. Vorch-streamer: Extending human audio-visual generation to real-time long-form streaming. *arXiv preprint arXiv:2608.05663*, 2026.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *CVPR*, pp. 2568–2577. IEEE, 2025.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jiahao Wang, Xu He, Zhenchao Tang, Meng Chu, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating. *arXiv preprint arXiv:2606.21661*, 2026.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pp. 21807–21818, 2024.

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 *CVPR*, pp. 28405–28415. IEEE, 2025.

Haoran Li, Fredreic Li, Shichen Ma, Jie Huang, Yijun Liu, Jiaqi Shi, and Yanwen Ma. Joyai-echo: Pushing the frontier of long audio-visual generation. 2026a.

Liyang Li, Wen Wang, Canyu Zhao, Tianjian Feng, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Mmcontrol: Unified multi-modal control for joint audio-video generation. *arXiv preprint arXiv:2604.19679*, 2026b.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. Omnicustom: Sync audio-video customization via joint audio-video generation model. *arXiv preprint arXiv:2602.12304*, 2026c.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *ICCV*, pp. 13848–13858. IEEE, 2025.

Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *ICLR*, volume 2026, pp. 139160–139194, 2026a.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Scofield Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *ICLR*, volume 2026, pp. 150592–150618, 2026b.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, et al. Mavin: Multi-shot audio-visual generation with narrative control. *arXiv preprint arXiv:2606.29473*, 2026c.

Yichen Liu, Quanwei Zhang, Haozhe Wang, Donghao Zhou, Xiaojie Li, Yang Shi, Jiaming Liu, Ruihua Huang, Yingtian Zou, and Daquan Zhou. The missing temporal link: Temporal context routing for script-driven audio-video generation. *arXiv preprint arXiv:2609.02367*, 2026d.

Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *ECCV*, pp. 468–485. Springer, 2024.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. In *CVPR*, pp. 461–471, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, pp. 10219–10228. IEEE, 2023.

Mingzhen Sun, Weining Wang, Yanyuan Qiao, Jiahui Sun, Zihan Qin, Longteng Guo, Xinxin Zhu, and Jing Liu. Mm-ldm: Multi-modal latent diffusion model for sounding video generation. In *ACM MM*, pp. 10853–10861, 2024.

Linrui Tian, Qi Wang, and Bang Zhang. Streamchar: Long-horizon streaming character audio-video generation with decoupled orchestration. *arXiv preprint arXiv:2605.25659*, 2026.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024.

Xuan Wei, Longbin Ji, Guan Wang, Xiangrui Liu, Zhenyu Zhang, Shuohuan Wang, Yu Sun, and Qingqi Hong. Memento: Reconstruct to remember for consistent long video generation. *arXiv preprint arXiv:2606.14667*, 2026.

Ziyi Wu, Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Yuwei Fang, Varnith Chordia, Igor Gilitschenski, and Sergey Tulyakov. Mind the time: Temporally-controlled multi-event video generation. In *CVPR*, pp. 23989–24000. IEEE, 2025.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, pp. 7151–7161. IEEE, 2024.

Xin Yan, Yuxuan Cai, Qiuyue Wang, Yuan Zhou, Wenhao Huang, and Huan Yang. Long video diffusion generation with segmented cross-attention and content-rich video data curation. In *CVPR*, pp. 3184–3194. IEEE, 2025.

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. In *CVPR*, pp. 1950–1960, 2026a.

Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025.

Lisai Zhang, Yidi Wu, Qi Liu, Xin Ma, Yang Ding, Gang Yue, Siqian Yang, Jingyuan Chen, Lin Ma, and Yaohui Wang. Vorch-director: Interactive world story model via noise-aware error rectification. *arXiv preprint arXiv:2608.05776*, 2026b.

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 *ICLR*, volume 2025, pp. 50060–50090, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 37:110315–110340, 2024.

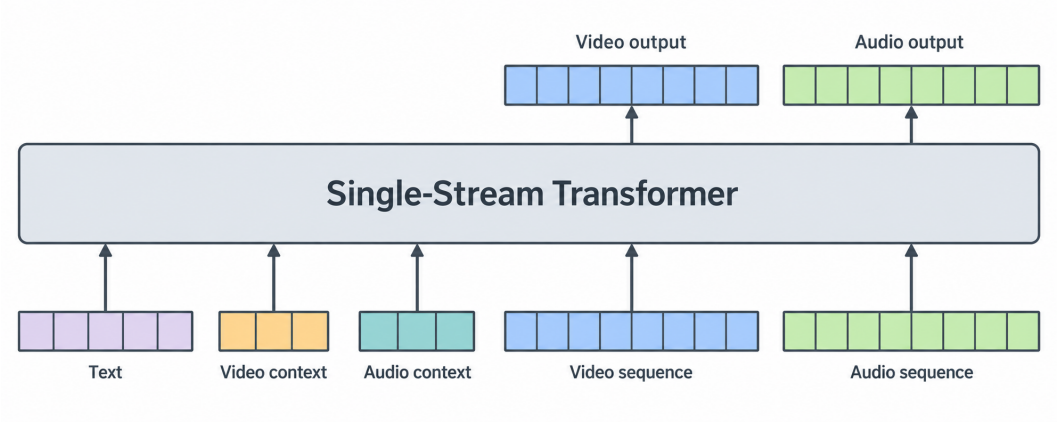

Figure 5: Overview of the single-stream audio-video transformer. Face and voice references occupy the video and audio context groups, respectively. The video context group also includes sparse visual memory for multi-shot continuation. For continuation, the initial tokens within each audio-video sequence are replaced with the corresponding history latents. The remaining positions contain noisy target latents, for which the transformer predicts video and audio denoising velocities.

A Model Implementation

This section details the encoders, token organization, and positional coordinates used by the shared audio-video generator.

We first train an in-house text-to-audio-video backbone with a single-stream Transformer architecture similar to Z-Image (Cai et al., 2025). The pretrained backbone initializes the subsequent three-stage training procedure.

A.1 Encoders and Reference Preparation

The frozen visual encoder is the Wan2.2 video VAE Wan et al. (2025), with temporal compression by 4, spatial compression by 16×16, and 48 latent channels. Our separately trained in-house stereo audio VAE remains frozen and operates at 24 kHz with a hop size of 2048, giving approximately 11.72 latent steps per second. The stereo channels are combined along the latent channel dimension. The frozen Qwen3-VL-8B-Instruct encoder Bai et al. (2025) supplies penultimate hidden states, which are normalized and projected to the transformer width. The maximum context length is 2048 tokens. Face crops are background-matted, resized, and padded with white to a 384×384 canvas before encoding. Videos are sampled at a target frame rate of $12 - 15$ FPS and organized into aspect-ratio buckets. Reference-face placeholders `@CTX_IMG_k` are grouped in a prefix before the complete script and replaced by the corresponding face images. No image placeholders are inserted into individual shot descriptions or dialogue. Each face is provided to both the multimodal prompt encoder and the video VAE. Sparse memory frames enter only through the visual VAE and are not supplied as images to Qwen3-VL.

The script describes characters and speakers in natural language. For multiple speakers, face and voice references follow the same order, with the first audible speaker's conditions placed first. This reference order is fixed across generation steps. The model learns the correspondence between reference conditions and script characters from the script and audio-video context. Reference indices denote positions in the condition sequence rather than explicit character labels in the script.

A.2 Transformer Layout and Attention

The shared DiT has 48 layers, hidden width 4608, 36 attention heads, and head dimension 128. Each modality refiner has two layers, and the three RoPE axes use dimensions 32, 48, and 48. Modality-specific projections and refiners map video, audio, and encoded script features to the shared width. The Transformer has separate video and audio output heads.

Figure 5 illustrates the input token groups in their concatenation order and the corresponding outputs. Face and voice references enter through the video and audio context groups, respectively. For multi-shot continuation, sparse visual memory also enters the video context group.

Let F_k and U_k denote encoded face and voice references, K the number of reference slots, and C_i^{enc} the multimodal script-encoder output. Let H_i^v, H_i^a denote latent history tokens and S_i the concatenated tokens of sparse visual memory $\mathcal{S}_i$. After projection, we form the context groups and temporal sequences as

$$\begin{aligned} R_i^v &= [F_1, \ldots, F_K, S_i], \qquad R_i^a = [U_1, \ldots, U_K], \\ Y_i^v &= [H_i^v, X_i^v], \qquad Y_i^a = [H_i^a, X_i^a], \\ Z_i &= [C_i^{\text{enc}}, R_i^v, R_i^a, Y_i^v, Y_i^a], \end{aligned} \tag{4}$$

where R_i^v, R_i^a are the video and audio context groups, and Y_i^v, Y_i^a are the corresponding temporal sequences. For continuation, history latents replace the initial tokens within each temporal sequence, while X_i^v, X_i^a contain noisy latents at the remaining target positions. The two modalities can have different history lengths. Absent reference and memory groups are omitted, while a face-only or voice-only reference retains its assigned slot. The transformer predicts denoising velocities only for the audio-video target tokens.

Let $\mathcal{I}_i^c$ contain the indices of reference, memory, and history tokens, and let $\mathcal{I}_i^j$ contain the indices of target and script tokens. These sets identify token roles within the input sequence. For valid query and key positions p and q, the additive attention mask is

$$M_{pq} = \begin{cases} -\infty, & p \in \mathcal{I}_i^c,\ q \in \mathcal{I}_i^j, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Reference, memory, and history tokens attend only to tokens in $\mathcal{I}_i^c$. Target and script tokens can attend to all valid tokens, including one another. The mask does not impose temporal causality or restrict target tokens to a selected shot description.

A.3 Reference Slots and Positional Coordinates

We assign rotary positional coordinates independently of the order of tokens in the input sequence. Video tokens use three-axis temporal and spatial coordinates, while script and temporal audio tokens repeat their sequence or temporal coordinate on all three axes. Let p_{end} be the last script position, K the number of reference slots, and M the number of sparse memory frames. Context slots use fixed positional spacing,

$$b_j = p_{\text{end}} + j\Delta_{\text{ctx}}, \qquad j = 1, \ldots, K + M. \tag{6}$$

For paired reference k, face tokens use (b_k, h, w) and every voice-reference token uses (b_k, b_k, b_k). The paired conditions remain separate token groups but share the temporal slot. A reference with only one modality still occupies one slot. Memory frames use the subsequent visual slots without corresponding audio reference groups.

The temporal sequence of video history and target tokens starts after the context slots with a positional gap Δ_{seq}. The audio timeline is shifted to align the target starts as described below. Both gaps provide positional separation without inserting tokens.

A.4 Temporal Alignment

We use a reference frame rate of 15 FPS to define the shared latent-time coordinate scale. For audio latent rate $r_a = 24000/2048$, the initial audio step is calibrated using a 93-frame clip at this reference rate,

$$\alpha_a = \frac{23}{r_a(93/15) - 1}. \tag{7}$$

The numerator is the temporal coordinate span of its 24 video latents. This reference clip sets the coordinate scale rather than the duration of every generated segment. For video frame rate f, the video step is $\alpha_v = 15/f$, where 15 FPS is the reference rate used to set the coordinate scale. Let

b be the video-history origin and n_v, n_a the video and audio history latent counts. The temporal coordinates are

$$\begin{aligned} p_v(j) &= b + j\alpha_v, \\ p_a(\ell) &= b + n_v\alpha_v + (\ell - n_a)\alpha_a. \end{aligned} \tag{8}$$

Indices start at zero, so the first target tokens occur at n_v and n_a. Their coordinates coincide at $p_v(n_v) = p_a(n_a) = b + n_v\alpha_v$. History and target share a continuous coordinate timeline within each modality, without an additional gap at the generation boundary.

B Training Details

B.1 Training Data

We combine four complementary data sources comprising approximately 1 million short single-person single-shot clips, 1 million short two-person single-shot clips, 0.5 million long single-person single-shot sequences, and 0.3 million multi-shot sequences. The short clips provide supervision for face and voice conditioning, while the long single-shot and multi-shot sequences support continuation and shot transitions. The stage-specific use of these sources is described in Sec. B.2.

Our training data are obtained from online sources under applicable licenses or explicit authorization for the intended research use.

B.2 Three-Stage Training

Starting from the T2AV backbone, we train RaAVGen in three successive stages on 32 GPUs, with each stage initialized from the preceding checkpoint. The first stage runs for 10,000 optimization steps on short single-person and two-person clips to learn face and voice conditioning. The second stage runs for 20,000 steps on short multi-shot sequences with timestamped scripts to learn shot transitions and temporal control. The third stage runs for 10,000 steps on predominantly long single-shot sequences supplemented by long multi-shot data to learn history-conditioned continuation and sparse visual memory conditioning. The first two stages use short sequences without temporal history, while retaining the script and optional reference conditions.

B.3 Continuation Sampling and Script Augmentation

In the third stage, contiguous subsequences are sampled from ground-truth audio-video recordings. For continuation examples, the preceding portion supplies history and the following portion supplies the target. By default, the history consists of 4 video latent frames and the most recent 6 seconds of audio, both ending at the target start. Initial clips have no temporal history.

Pseudo-long script augmentation. We extend the textual context of existing recordings to train script localization without matching full-length audio-video data. For continuation examples with audio history, we construct $\widetilde{C} = C^- \oplus C \oplus C^+$, where $\oplus$ denotes concatenation. For single-shot data, C^- and C^+ contain additional speech text. For multi-shot data, they contain additional shot descriptions with timestamps. In both cases, the original audio-video history and target remain unchanged. The model learns to locate the relevant content within the extended script using its audio history and continue with the corresponding target.

Fig. 6 illustrates the single-shot case. Additional speech text is prepended to the script of an 18-second recording, placing the original content later in the extended script. History and target are still sampled from the original recording. The illustrated times indicate a conceptual offset only, since single-shot scripts contain no timestamps.

For multi-shot data, the augmentation also adjusts the script timestamps, as illustrated in Fig. 7. If the prepended shots span Δ seconds, each original shot interval $[s_j, e_j)$ becomes $[s_j + \Delta, e_j + \Delta)$. The sampled history and target retain their original audio-video content and correspond to the shifted positions in the extended script. Appended shot descriptions extend the script beyond the original recording. Neither the preceding nor the following additions introduce audio-video supervision.

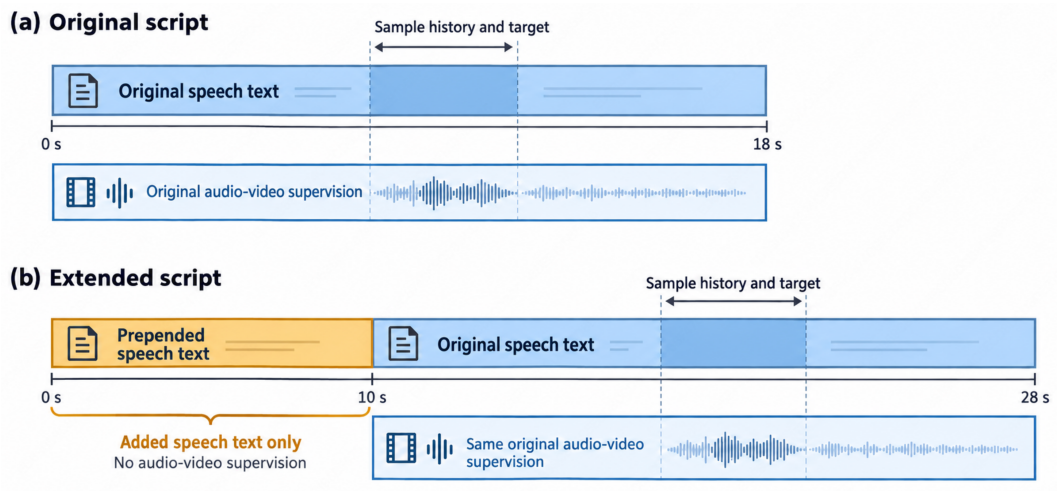

Figure 6: Pseudo-long script augmentation for single-shot data. (a) History and target clips are sampled from an 18-second recording paired with its original speech text. (b) Additional speech text is prepended to the script, placing the original content later in the extended script, illustrated here from 10 to 28 seconds. The audio-video history and target remain unchanged, and the added text provides no additional audio-video supervision. The illustrated time offset is conceptual, since single-shot scripts contain no timestamps.

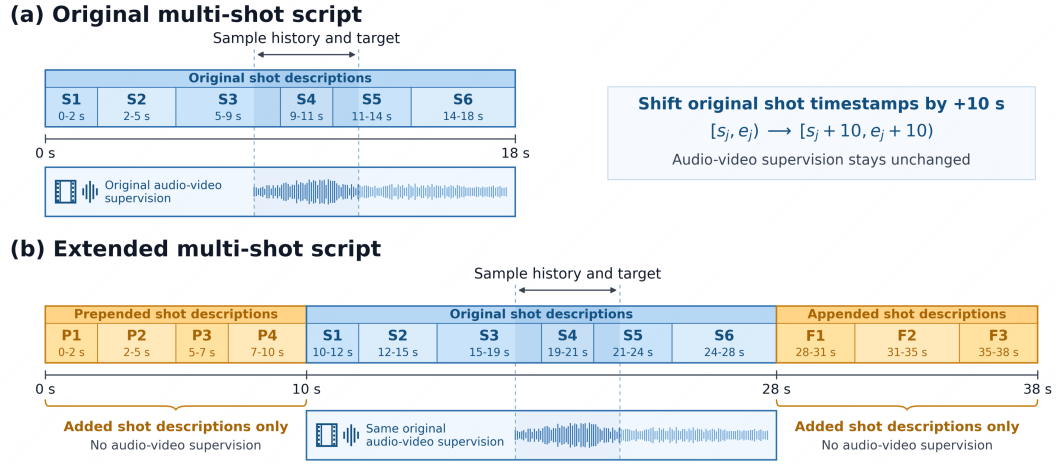

Figure 7: Pseudo-long script augmentation for multi-shot data. (a) The original script contains six shots spanning 18 seconds and is paired with an audio-video recording. (b) Four preceding shots and three following shots each add 10 seconds, extending the script to 38 seconds. The original shot descriptions, shown in blue, shift from $[0, 18)$ to $[10, 28)$. The audio-video recording and sampled history and target remain unchanged, with their corresponding script positions shifted by 10 seconds. The added shot descriptions, shown in orange, provide text only.

B.4 Reference Dropout

On datasets supporting reference dropout, the probabilities of no references, face only, voice only, and both are 0.1, 0.1, 0.1, and 0.7.

B.5 Joint Flow Matching

For modality $m \in \{v, a\}$, let x^m be the clean target and ϵ^m independent Gaussian noise. We sample $t \in (0, 1)$ from a logit-normal distribution and use modality-specific noise levels,

$$\begin{aligned} t_a &= t, \qquad t_v = \frac{16t}{1 + 15t}, \\ x^m_{t_m} &= (1 - t_m)x^m + t_m\epsilon^m, \\ u^{m,*} &= x^m - \epsilon^m. \end{aligned} \tag{9}$$

The model predicts u^m_θ to match the denoising velocity $u^{m,*}$ and receives a time embedding proportional to $1 - t_m$. This velocity convention points from noise toward data. Both modalities are predicted jointly, with loss evaluated only on valid target coordinates,

$$\begin{aligned} \mathcal{L}_m &= \mathbb{E}\left[\frac{1}{|\Omega_m|}\sum_{q\in\Omega_m}\left(u^m_\theta(q) - u^{m,*}(q)\right)^2\right], \\ \mathcal{L}_{\text{flow}} &= \lambda_v\mathcal{L}_v + \lambda_a\mathcal{L}_a. \end{aligned} \tag{10}$$

Here Ω_m excludes references, memories, history prefixes, and invalid target regions, and $\lambda_v = \lambda_a = 1$. Condition prefixes are normally kept clean, with a small optional perturbation to video history during training. They remain excluded from the target loss. The pretrained encoders are frozen, while the transformer and its input and output layers are optimized.

B.6 Optimization

For these three stages, we use a learning rate of 4×10^{-5}, weight decay of 0.01, and a second-moment coefficient of 0.95. The learning rate remains constant after 100 warmup steps, and the gradient norm is clipped to 1.0. Training uses bfloat16 precision and gradient checkpointing, with a per-device batch size of one and no gradient accumulation. Video and audio flow losses receive equal weights.

C Inference Details

C.1 Segment-by-Segment Generation

Each segment starts from noisy video and audio target latents. The first segment is conditioned on the complete script and optional face and voice references, without temporal history. Subsequent segments additionally receive asymmetric audio-video history and, for multi-shot narratives, sparse visual memory. Face and voice references remain fixed throughout the sequence.

C.2 Modality-Aware Guidance

We evaluate the joint audio-video branch and corresponding single-modality branches. Let $u^{+,va}_m$ denote the positive-condition prediction from the joint branch, and let $u^{+,m}_m$ and $u^{-,m}_m$ denote positive and negative predictions from the single-modality branch. We combine them as

$$\begin{aligned} \widehat{u}_m = u^{-,m}_m &+ g_m\left(u^{+,m}_m - u^{-,m}_m\right) \\ &+ d_m\left(u^{+,va}_m - u^{+,m}_m\right). \end{aligned} \tag{11}$$

The coefficient g_m controls positive-condition guidance, and d_m controls the contribution of the other generated modality. Positive single-modality branches retain the references and temporal conditions relevant to that modality. In particular, the audio-only branch retains voice references and audio history. The negative branch removes reference and sparse-memory contexts in addition to using negative text conditioning.

D Evaluation Details

D.1 Evaluation Protocol

We evaluate all methods on the same story specifications within each comparison and use the same face and voice references whenever supported. Input formatting follows each method's supported interface while preserving the intended dialogue, visual events, and reference conditions. For baselines that accept only local prompts, each local prompt is derived from the corresponding portion of the complete script and formatted according to the prompt template provided on the method's official website. Metrics are computed from generated outputs using a shared evaluation pipeline.

D.2 Evaluation Metrics

We evaluate generated videos along complementary dimensions of content accuracy, reference fidelity, long-range consistency, temporal structure, and perceptual quality. For speech content and audio-visual alignment, we report word error rate (WER) using Whisper large-v3 Radford et al. (2023), the Time Alignment Metric for Speech (TAMS) Liu et al. (2026c), and SyncNet confidence/distance (Sync-C/D) Chung & Zisserman (2016). We assess perceptual visual quality using the VBench Huang et al. (2024) aesthetic and imaging-quality scores (AES/IQ), and audio quality using the Content Usefulness and Production Quality scores (CU/PQ) of Audiobox Aesthetics Tjandra et al. (2025). Reference-condition fidelity is measured by face identity similarity (ID-Sim) and voice timbre similarity (T-Sim), computed using ArcFace Deng et al. (2019) and WavLM Chen et al. (2022) speaker embeddings, respectively. For long-range consistency, we report character and background inter-shot consistency (CISC/BISC) Liu et al. (2026c), together with adapted cross-shot identity and timbre consistency scores (X-ID-Sim/X-T-Sim). Shot structure is evaluated using Shot Boundary MAE (SB-MAE), computed over matched shot pairs following TCR Liu et al. (2026d). We additionally report shot-boundary F1 at a tolerance of $0.5\,\mathrm{s}$ (SB-F1@0.5s). Predicted and reference boundaries are matched one-to-one within this tolerance: matched pairs count as true positives, while unmatched predicted and reference boundaries count as false positives and false negatives, respectively. We additionally evaluate segment continuity, identity and voice consistency, and prompt following through the user study in Sec. E.1.

TAMS. We evaluate speech timing using TAMS, following MAVIN Liu et al. (2026c). It measures temporal intersection-over-union between target speech intervals and detected speech activity, with higher scores indicating better temporal alignment.

Cross-shot identity consistency. X-ID-Sim measures facial identity consistency across shots. Using fixed character assignments, we extract ArcFace embeddings Deng et al. (2019) from valid face observations, average and normalize them within each shot, and compute cosine similarity between different shots containing the same character.

Cross-shot timbre consistency. X-T-Sim follows the same procedure using WavLM speaker-verification embeddings Chen et al. (2022) from valid speech turns. For both metrics, similarities are averaged over valid shot pairs within each character, then equally across valid characters within each story, and finally equally across valid stories. Invalid observations are excluded, and the resulting scores summarize consistency over valid shot pairs. These metrics measure cross-shot self-consistency rather than reference fidelity, and are therefore reported together with ID-Sim and T-Sim.

E Additional Experiments

We report a comparative user study and ablations of audio history length, video history length, and pseudo-long script augmentation. Each user study involves 30 participants and 30 randomly selected cases. Participants assign scores from 1 to 5, with higher scores indicating better performance.

Table 2: User study results on 30 examples evaluated by 30 participants. Scores range from 1 to 5, with higher values indicating better performance. We report mean ratings for continuity across segments, identity and voice consistency, and prompt following.

Method	Segment continuity ↑	Identity and voice consistency ↑	Prompt following ↑
Ovi	3.2	2.9	3.0
DreamID-Omni	3.0	2.6	3.1
TCR	3.3	3.2	3.4
H3	3.9	**3.8**	3.6
JoyAI-Echo	3.1	3.1	2.8
ours	**4.1**	**3.8**	**3.9**

Table 3: Effect of audio history duration on speech accuracy. Video history length and all other settings are held fixed. The default configuration uses 6 seconds of audio history.

Audio history duration	WER ↓
10 s	2.66
8 s	2.71
6 s (default)	**2.64**
4 s	3.15
2 s	4.08

E.1 Comparative User Study

Participants assess the generated audio-video sequences along three dimensions. Segment continuity measures how naturally speech and visual content continue across generation boundaries. Identity and voice consistency measures whether characters retain consistent appearance and vocal characteristics throughout the sequence. Prompt following measures how closely the generated content follows the input script. We average the ratings for each dimension and report the results in Table 2. RaAVGen receives the highest mean ratings for segment continuity and prompt following, while matching H3 on long-form identity and voice consistency.

E.2 Effect of Audio History Length

We evaluate audio history durations of 2, 4, 6, 8, and 10 seconds while fixing video history at 4 latent frames and keeping all other settings unchanged. Our default configuration uses 6 seconds of audio history. As shown in Table 3, WER remains similar when at least 6 seconds of audio history are available. With shorter histories, we observe more word errors near segment boundaries, suggesting difficulty in identifying where the preceding speech ends within the script. These results support using 6 seconds of audio history to provide sufficient speech context.

E.3 Effect of Video History Length

We vary the video history length across 2, 4, and 8 latent frames while fixing the audio history duration at 6 seconds. We report the number of video conditioning tokens and the average generation time per segment under identical hardware and sampling settings. For this ablation, each segment generates 97 new target frames at a resolution of 448×896. Historical conditioning frames are excluded from this count, and the target frame count remains fixed across all three configurations. We also conduct a user study in which participants rate the smoothness of transitions across segment boundaries using the five-point scale described above. Only boundaries within the same shot are evaluated to avoid penalizing intended shot changes.

As shown in Table 4, increasing video history from 2 to 4 latent frames improves the mean user study score from 2.7 to 4.3, while generation time increases from 7.13 to 7.60 minutes per segment. Using 8 latent frames provides a smaller gain to 4.5, but doubles the number of video conditioning

Table 4: Effect of video history length on conditioning token count, generation time, and visual continuity. Each segment generates 97 new target frames at 448×896 resolution, excluding historical conditioning frames.

Video history	Video history tokens	Time per segment (min) ↓	User study ↑
2 latent frames	3136	7.13	2.7
4 latent frames (default)	6272	7.6	4.3
8 latent frames	12544	8.75	**4.5**

Table 5: Effect of pseudo-long script augmentation on speech accuracy across generation intervals. Both variants use the same source recordings of at most 20 seconds. Augmentation extends the script context to at most 45 seconds while retaining the original audio-video supervision. Lower WER is better.

Training variant	WER (%) ↓		
	$[0, 8)$ s	$[8, 16)$ s	$[16, 24)$ s
Without augmentation	2.32	2.37	5.85
With augmentation	2.35	2.48	2.53

tokens from 6,272 to 12,544 and increases generation time to 8.75 minutes. We therefore use 4 latent frames by default to balance visual continuity and computational cost.

E.4 Effect of Pseudo-Long Script Augmentation

We examine the effect of pseudo-long script augmentation on speech accuracy at different stages of generation. For this controlled comparison, all recordings used for long-sequence training are limited to 20 seconds, with approximately 90% lasting between 16 and 20 seconds. We compare two variants trained on the same source recordings with and without script augmentation, keeping all other training and inference settings fixed. The augmented variant prepends and appends additional script content with randomly sampled lengths, extending the total script span to at most 45 seconds. Following Sec. B.3, the additions consist of speech text for single-shot data and timestamped shot descriptions for multi-shot data. The original audio-video history and target remain unchanged, so the augmentation does not require longer recordings.

Both variants are evaluated on the same test scripts. We report WER separately for the intervals $[0, 8)$, $[8, 16)$, and $[16, 24)$ seconds, measured from the start of generation. As shown in Table 5, the two variants exhibit comparable WER in the earlier intervals. Without augmentation, WER increases markedly in the final interval, whereas it remains relatively stable with augmentation. These results suggest that pseudo-long script augmentation helps maintain speech accuracy during extended generation despite the limited duration of training recordings.

F Limitations and Future Work

Error accumulation is a shared challenge in autoregressive generation. As each segment depends on earlier outputs, speech errors or visual drift may propagate during extended generation. RaAVGen follows this generation paradigm and remains subject to this limitation. The quality of complex actions also depends on the capabilities of the underlying audio-video backbone. Our framework does not specifically address the motion modeling needed for intricate actions and interactions.

This work focuses on locating narrative progress within a complete script and continuing generation across segment boundaries. Addressing error accumulation and improving complex motion synthesis remain complementary research directions. Future work will explore stronger backbones and mechanisms for correcting continuation errors, while extending the framework to conversations with more speakers and narratives involving more complex scenes.