SCENEGEN: GENERATING MULTI-SHOT SCENES WITH TIMED SPEECH AND SOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in joint audio-video generation have enabled realistic scenes with synchronized speech and sound. However, fine-grained temporal control remains challenging for open models, especially when camera transitions, visual actions, and audio events exhibit distinct temporal patterns within the same scene. Global conditioning cannot define local timing, while shot-level conditioning restricts events to camera boundaries, disrupting cross-shot continuity. To address this limitation, we propose SceneGen, a framework for fine-grained temporal conditioning that models shot transitions and audio events separately. Event-Aware Attention independently encodes complete speech and sound events, activating them based on their timeline intervals. Shot-Aware RoPE explicitly represents camera transitions through shared temporal position shifts across the audio and visual modalities. To provide supervision consistent with these mechanisms, we construct SceneGen-Data, containing approximately 194K multi-shot scenes with raw audio, linked character identities across views, shot-level visual annotations, and independently timed audio events. Together, these designs allow the model to follow shot-specific instructions while preserving events across camera cuts. Experiments on SceneGen-Bench demonstrate improved speech timing, prompt following, and cross-shot character consistency, maintaining overall audiovisual quality. We will make our dataset and models publicly available.

1 INTRODUCTION

Audio-video generation has advanced rapidly in both industry and academia. Recent systems can generate videos with dialogue, music, and environmental sounds (ByteDance Seed, 2026; Alibaba Cloud, 2026), while open models further provide realistic motion and synchronized audio with publicly available weights (Lightricks, 2026; Low et al., 2025). These advances provide a strong foundation for controlling complex scenes with multiple shots, recurring characters, camera changes, and temporally structured audio events.

Recent work has extended generation from isolated clips to coherent shot sequences. Existing methods condition individual shots while maintaining visual continuity across transitions (Kara et al., 2025; Wang et al., 2025b), while related audio-video approaches introduce temporally localized descriptions for different parts of a scene (Liu et al., 2026). In parallel, datasets have progressed from clip-level captions toward dense temporal descriptions and linked audiovisual annotations (Chen et al., 2024; Xiong et al., 2024; Chen et al., 2026). However, fine-grained control remains challenging because visual instructions and audio events often follow different temporal schedules.

A camera cut may change the visible character, action, or viewpoint while an utterance continues across the transition. Binding all local instructions to shot boundaries can therefore interrupt ongoing audio events, whereas global conditioning does not explicitly specify the temporal scope of local instructions. Figure 1 illustrates this limitation. In panel A, narration begins later than requested, while in panel B, the model retains a wide facade view instead of following the requested camera descent. These examples motivate our central question: ***How can data and generation algorithms jointly give an open audio-video model fine-grained control within and across shots?***

To address this problem, we propose SceneGen, a fine-grained multi-shot audio-video generation framework that separates the temporal modeling of camera transitions from that of speech and sound events. Event-Aware Attention independently encodes each complete audio event and activates its

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.

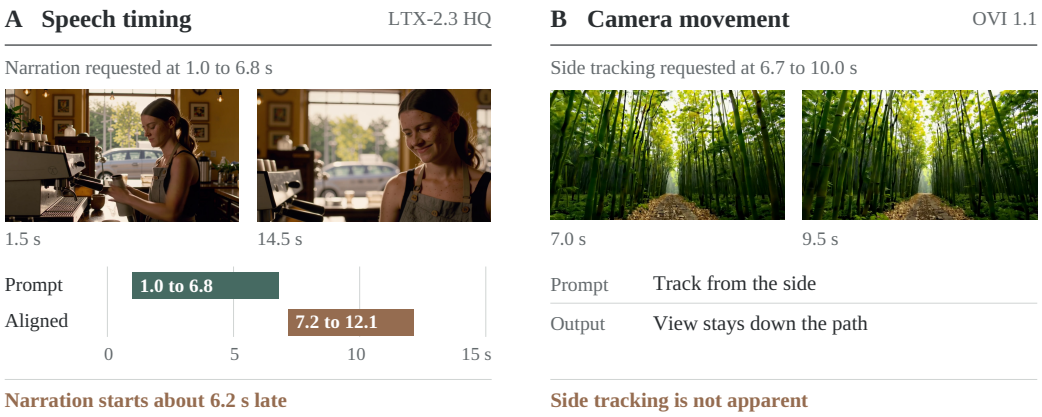


Figure 1: **Speech timing and camera control motivate fine-grained conditioning.** Panel A compares the requested speech timing with automatic alignment. Panel B requests a descent toward the blue door, but the generated video remains in a wide view instead of following the requested camera movement.

representation only during the corresponding interval on the original scene timeline. This allows speech and sound to remain active across camera cuts without being constrained by shot boundaries. For visual transitions, we use Shot-Aware RoPE to explicitly encode shot boundaries, enabling the model to perceive structural transitions between shots.

To provide supervision consistent with this design, we construct SceneGen-Data, a dataset of approximately 194K multi-shot scenes. Our pipeline preserves consecutive footage and original audio, links recurring character identities across views, and separately annotates visual actions and temporally localized audio events. Audio event boundaries remain independent of camera transitions, aligning the training supervision with the temporal structure used by SceneGen.

Experiments on SceneGen-Bench validate the effectiveness of this design. The proposed approach achieves improved speech timing, instruction following, and cross-shot character consistency, indicating that the model can better preserve both temporal alignment and semantic continuity across shot transitions. At the same time, these improvements are obtained without degrading the overall audiovisual quality, demonstrating that the enhanced temporal modeling does not compromise general generation quality.

In summary, our contributions are threefold:

- We construct SceneGen-Data, a multi-shot audio-video dataset with approximately 194K scenes and fine-grained annotations that separately describe shot-level visual actions, temporally localized speech and sound events, and recurring character identities across shots.
- We propose SceneGen, a fine-grained temporal conditioning framework that combines Event-Aware Attention with Shot-Aware RoPE to separately model independently timed audio events and camera transitions.
- We enable fine-grained controllable multi-shot audio-video generation while preserving overall audiovisual quality, achieving improved speech timing, instruction following, and cross-shot character consistency on SceneGen-Bench.

2 METHODOLOGY

We propose SceneGen, a fine-grained audiovisual generation framework built on LTX-2.3 (Lightricks, 2026), together with SceneGen-Data, a multi-shot audiovisual dataset with linked character identities and temporally structured visual and audio annotations. SceneGen consists of Shot-Aware RoPE for modeling camera transitions and Event-Aware Attention for aligning event descriptions with their temporal intervals. The dataset construction and the two model components are described in the following sections.

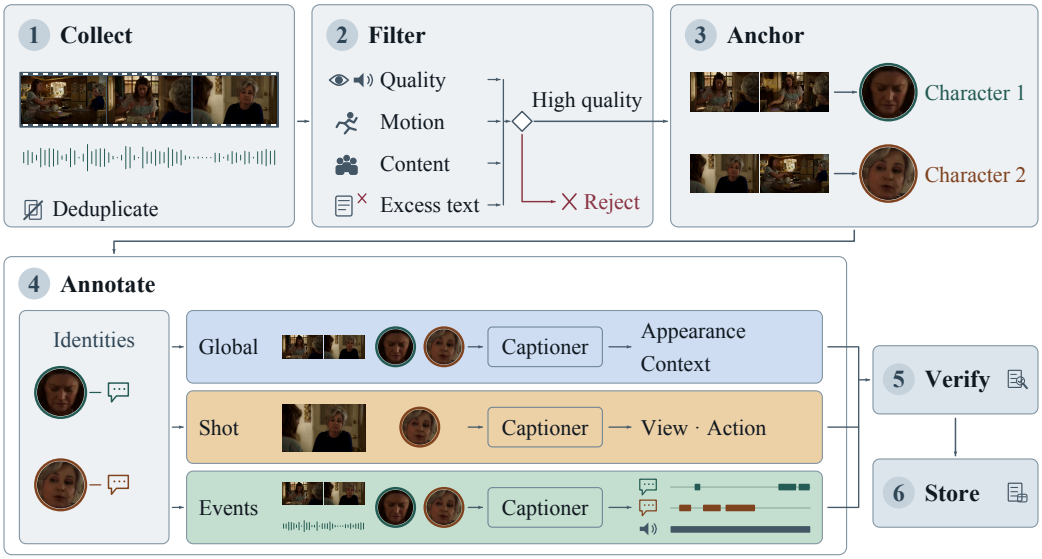


Figure 2: **Construction of SceneGen-Data.** Continuous multi-shot scenes are first collected and filtered for audiovisual quality. Character identities are then linked across shots, followed by scene-level, shot-level, and event-level annotation.

2.1 Dataset Construction

Fine-grained audiovisual control requires training data that preserves continuous multi-shot content while providing annotations at different temporal scales. We therefore construct SceneGen-Data through four stages: scene collection, quality filtering, character anchoring, and fine-grained annotation. The resulting dataset contains approximately 194K multi-shot scenes with original audio, linked character identities, shot-level visual descriptions, and independently timed audio events.

Scene collection. We collect videos from Koala-36M (Wang et al., 2025c), Panda-70M (Chen et al., 2024), and Internet sources. After source-level deduplication, all videos are standardized to 720p at 24 fps. We detect shot boundaries using TransNetV2 (Soucek & Lokoc, 2024) and group consecutive shots into scenes between 5 and 30 seconds, with durations centered around 15 seconds. Each scene retains contiguous footage and its original audio. We further trim 0.5 seconds from both ends and decrease the error that could be caused by the TransNetV2.

Quality filtering. The collected scenes may contain low-quality media, static content, editing artifacts, or overly crowded compositions. We evaluate visual and audio quality using DOVER++ (Wu et al., 2023) and Audiobox Aesthetics (Tjandra et al., 2025), and use RAFT (Teed & Deng, 2021) to identify static or slideshow-like content. We also reject scenes with persistent black bars or more than five high-confidence person detections in most sampled frames. RapidOCR (RapidAI Team, 2021) detects subtitles and logos, and STTN (Zeng et al., 2020) removes these regions before the original aligned audio is remuxed with the processed video.

Character anchoring. After filtering, character identities must remain consistent across changes in viewpoint, scale, and composition. We detect faces and body poses in sampled frames, match face embeddings across observations, and progressively merge similar identity clusters. For each identity, reference crops are selected according to detection confidence, face size, and frontal visibility. These references are then used to associate the same character across different shots during annotation.

Fine-grained annotation. With continuous scenes and linked identities established, we annotate each scene at three temporal levels: scene-level appearance, shot-level visual content, and time-aligned audio events. WhisperX (Bain et al., 2023) transcribes speech in the corpus languages, DiariZen (Han et al., 2025a) performs speaker diarization, and Qwen3-Omni (Xu et al., 2025) associates voices with character identities.

Each annotation type is generated using context matched to its temporal scope. Scene-level descriptions use frames and identities from the complete scene. Shot-level captions use local frames

Table 1: **Annotation scope of SceneGen-Data and related datasets.** N/A indicates a field not established by the cited source. Sound timing refers to intervals defined independently of shot boundaries.

Dataset	Visual scope	Characters / speakers	Speech timing	Sound timing
Panda-70M (2024)	Clip	N/A	N/A	N/A
FineVideo (2024)	Narrative	Roster	Timed ASR	N/A
Cine250K (2025)	Shot	N/A	N/A	N/A
SpeakerVid-5M (2025)	Person	Linked	ASR + turns	N/A
CineDance-1M (2026)	Shot	Linked	Sentence	Shot
SceneGen-Data	Shot	Linked	Utterance	**Event**

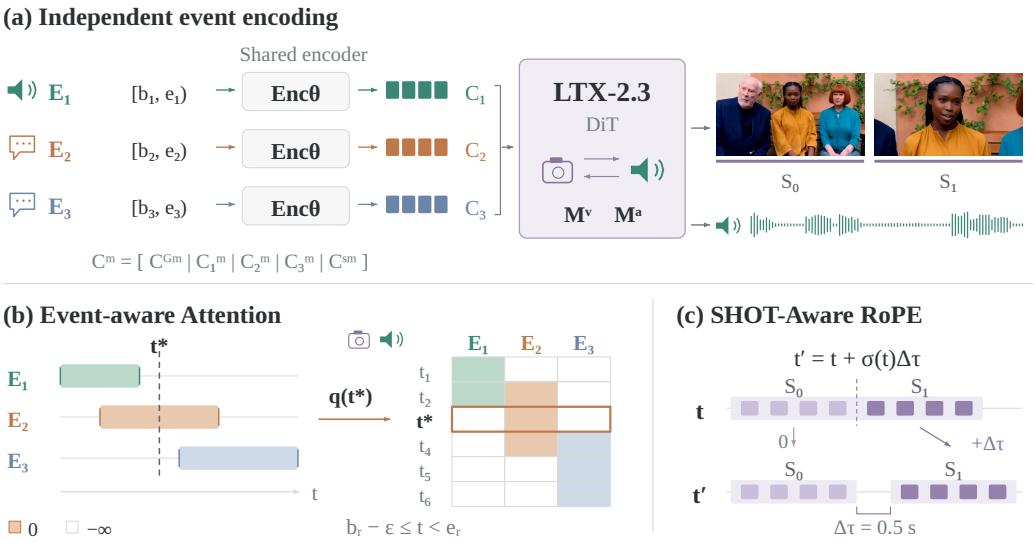


Figure 3: **Shot-Aware RoPE and Event-Aware Attention in SceneGen.** **(a)** Events are encoded independently using shared weights. **(b)** Filled cells indicate active events, and the highlighted query accesses E_2. **(c)** Position offsets represent camera cuts in both modalities, while event selection follows the original scene timeline.

and the identities present in the corresponding shot. Audio events use the audio track together with visual segments aligned to individual utterances. This process produces descriptions of visual actions, speech, vocal attributes, and non-speech sounds while retaining their corresponding temporal ranges. Annotation consistency is verified before storage.

Comparison with existing datasets. Table 1 compares SceneGen-Data with related datasets in terms of visual granularity, identity association, speech timing, and sound timing. Existing datasets provide different combinations of shot-level descriptions, linked identities, and speech alignment, while sound events are either not explicitly localized or tied to shot boundaries. Such representations do not fully capture acoustic temporal structure, since an event may occupy only part of a shot or persist across multiple camera cuts. SceneGen-Data instead represents visual content at the shot level while assigning speech and sound their own temporal intervals, allowing acoustic events to remain independent of visual transitions.

These annotations expose two distinct temporal structures for model conditioning. Shot boundaries capture changes in visual composition, whereas event intervals specify the active duration of speech and sound on the continuous scene timeline. Accordingly, SceneGen uses Shot-Aware RoPE to encode visual transitions and Event-Aware Attention to restrict event descriptions to their corresponding temporal intervals.

2.2 SHOT-AWARE ROPE

Temporal RoPE does not explicitly represent the discontinuity introduced by camera cuts. Following previous work (Wang et al., 2025b), we incorporate shot boundaries by adding a shot-dependent offset to the temporal coordinate.

Let t denote the original time in seconds and $\sigma(t)$ the zero-based shot index. For both modalities, we define

$$t' = t + \sigma(t)\Delta\tau, \qquad \Delta\tau = 0.5 \text{ s}. \tag{1}$$

This formulation preserves relative temporal distances within each shot while separating consecutive shots in the positional space. Spatial coordinates and rotary frequencies remain unchanged. For interval-valued positions, both endpoints are shifted according to the shot index of the interval center.

The offset is applied only to positional encoding. Event intervals remain on the original scene timeline, and audiovisual conditioning retains the original temporal coordinates to preserve the timing of the conditioning modality.

2.3 EVENT-AWARE ATTENTION

While Shot-Aware RoPE represents visual discontinuities across camera cuts, event descriptions require a different form of temporal control because their valid intervals are not determined by shot boundaries. Previous work often organizes audiovisual conditions using global, shot-level, and role-level structures (Liu et al., 2026), where semantic information is associated with visual or identity-based scopes. Such organization is less suited to events whose temporal extents may begin or end within a shot or continue across several shots. To address this mismatch, we introduce Event-Aware Attention, which represents each event as an independent semantic unit and determines its accessibility from its temporal interval. This decouples event timing from shot structure and restricts each media token to the event descriptions that are active at its time.

Aligning events with time. Let G, S_i, and E_r denote the text-key sets for global context, shot i, and event r. Event r occupies the interval $[b_r, e_r)$, and each media query is associated with its original token-center time t. To account for latent temporal discretization, we use an onset tolerance of $\epsilon = 0.02$ seconds. The accessible text keys for video and audio queries are defined as

$$\begin{aligned} \mathcal{I}^v(t) &= G \cup S_{\sigma(t)} \cup \bigcup_{r:b_r-\epsilon\le t<e_r} E_r, \\ \mathcal{I}^a(t) &= G \cup \bigcup_{r:b_r-\epsilon\le t<e_r} E_r. \end{aligned} \tag{2}$$

Both modalities receive global context. Video additionally receives the caption of the current shot because visual actions and camera motion are defined at the shot level. Event descriptions are accessible only while their intervals are active. When multiple events overlap, they remain jointly accessible until their respective endpoints. A camera cut therefore changes the active shot caption without affecting an event that continues on the original scene timeline.

Enforcing temporal alignment. We implement this access rule using an additive mask in media-to-text cross-attention. For a query q from modality $m \in \{v, a\}$ and a text key k, the mask is defined as

$$M^m_{q,k} = \begin{cases} 0, & k \in \mathcal{I}^m(t_q), \\ -\infty, & \text{otherwise}, \end{cases} \tag{3}$$

and the corresponding cross-attention (Vaswani et al., 2017) is

$$O^m = \text{softmax}\left(\frac{Q^m (K^m)^\top}{\sqrt{d_m}} + M^m\right) V^m. \tag{4}$$

Here Q^m, K^m, V^m, and d_m denote the modality-specific queries, text keys, values, and attention-head dimension. Global keys remain accessible to every query so that each attention row contains at least one valid key. Media self-attention and audiovisual cross-attention remain unchanged. The mask only restricts direct media-to-text access, preventing descriptions from affecting queries outside their intended temporal intervals.

Table 2: **Audio-video quality and fine-grained multi-shot control on SceneGen-Bench. Bold** and underlined values indicate the best and second-best available results for each metric. - denotes an unavailable metric.

Method	Video		Audio				Multi-shot Control				
	TA ↑	AES-V ↑	CU ↑	PQ ↑	WER ↓	Incorr. ↓	Desync ↓	PF ↑	CC ↑	CTE ↓	STE ↓
Multi-shot Video Generation											
HoloCine	19.16	0.528	-	-	-	-	-	4.08	4.02	**1.74**	-
MultiShotMaster	19.23	0.517	-	-	-	-	-	3.69	4.09	2.31	-
Synchronized Audio-Video Generation											
LTX-2.3	22.20	0.558	5.29	6.68	6.47	22.84	0.38	-	-	-	3.52
Ovi	18.86	**0.572**	5.51	5.92	4.58	14.81	0.43	-	-	-	2.67
OmniVAE T2AV	15.38	0.494	5.41	6.10	19.37	49.64	0.83	-	-	-	1.74
MOVA-720p	22.21	0.543	4.88	5.72	23.04	74.07	0.56	-	-	-	1.46
SceneGen	**23.44**	0.565	**5.56**	**6.77**	**0.79**	**14.32**	**0.28**	**4.19**	**4.31**	1.97	**0.23**

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Training Setup. We initialise SceneGen from the 22B LTX-2.3 model and train it at 720p and 24 fps for 3,800 steps on 64 H20 GPUs with a global batch size of 128. We use a peak learning rate of 10^{-5} with 500 warm-up steps followed by cosine annealing, a gradient clipping threshold of 1.0, BFloat16 precision, and ZeRO-3 optimisation. Length-budget sampling groups scenes of similar duration to reduce the training overhead caused by long scenes.

Training Dataset. As described in Sec. 2.1, our training data are collected from Panda-70M, Koala-36M, and internally collected Internet videos, resulting in approximately 197k scenes. Most scenes last 10 to 20 seconds, contain fewer than three people, and include fewer than five shots. All videos are standardised to 720p and 24 fps. Further details are provided in Appendix B.

Benchmark and Baselines. We construct SceneGen-Bench with 100 prompts covering cinematic, television, and vlog-style scenes. The benchmark includes dialogue, successive actions, multiple characters, shot transitions, and camera changes, with the goal of evaluating fine-grained control within coherent multi-shot sequences. We compare SceneGen against representative open-source baselines with publicly available weights, including HoloCine and MultiShotMaster for multi-shot video generation, and LTX-2.3, Ovi, OmniVAE T2AV, and MOVA-720p for synchronised audio-video generation.

Metrics. TA: text-video alignment measured with ViCLIP (Wang et al., 2024). **AES-V:** aesthetics measured with the LAION aesthetic predictor (LAION-AI, 2022; Radford et al., 2021). **CU and PQ:** audio content usefulness and production quality measured with Audiobox Aesthetics (Tjandra et al., 2025). **WER and Incorrect:** speech-content errors from Whisper large-v3 transcripts. **PF and CC:** GPT-based ratings of prompt following and cross-shot character consistency. **CTE:** shot-boundary timing error with TransNetV2 (Soucek & Lokoc, 2024). **STE:** average start and end boundary error for requested utterances. **Desync:** audio-visual speech offset measured with Synchformer (Iashin et al., 2024). Definitions and units are provided in Appendix C.3.

3.2 MAIN RESULTS

Table 2 evaluates three aspects that are considered separately: video quality, synchronized audio generation, and fine-grained multi-shot control. Existing multi-shot methods report strong results on shot structure but do not support the full set of audio measurements, while synchronized audio-video models provide audio generation capabilities without evaluation of shot-level semantic control. In contrast, SceneGen is evaluated across all three aspects and shows strong performance.

Fine-grained Speech Control. SceneGen shows a clear advantage in both speech accuracy and temporal control. It achieves a WER of 0.79, substantially lower than existing audio-video generation methods, while also obtaining the lowest incorrect-utterance rate of 14.32. The improvement is more pronounced in temporal precision, where SceneGen reduces STE to 0.23 compared with 1.46 for the strongest reported baseline. Meanwhile, its Desync score of 0.28 also improves over LTX-2.3 and Ovi, which achieve 0.38 and 0.43, respectively. These results indicate that SceneGen can generate the requested dialogue at the intended time while maintaining accurate audio-visual synchronization.

Fine-grained Multi-shot Control. Beyond speech control, SceneGen also performs strongly on shot-level semantic consistency. It achieves the highest PF and CC scores of 4.19 and 4.31, outperforming HoloCine and MultiShotMaster on both aspects. In particular, the improvement in CC indicates that stronger local instruction following does not compromise character consistency across changing shots. At the same time, SceneGen obtains a CTE of 1.97, remaining competitive with dedicated multi-shot methods such as HoloCine at 1.74 and outperforming MultiShotMaster at 2.31. This balance suggests that the model can preserve accurate shot organization while providing stronger control over the semantic content within each shot.

Audio-visual Quality. The gains in fine-grained control are achieved without degrading generation quality. SceneGen obtains the highest TA, CU, and PQ scores of 23.44, 5.56, and 6.77, respectively. Its AES-V score of 0.565 is close to the best of 0.572 from Ovi. Thus, the improved temporal and multi-shot controllability is accompanied by competitive visuals and stronger audio quality.

Overall, SceneGen provides a more balanced solution to multi-shot audio-video generation by jointly improving speech timing, audio-visual synchronization, shot-specific instruction following, and cross-shot consistency while preserving generation quality.

3.3 Qualitative Results

Figure 4 demonstrates that SceneGen can coordinate heterogeneous events within a shared temporal structure rather than binding all semantic instructions to shot boundaries. In the timed and three-speaker dialogue examples, speech is generated in the requested order and temporal intervals while maintaining the corresponding speaker identities as the visual focus changes. These examples show that the model can associate temporally localised dialogue with persistent characters even when multiple speakers appear in the same sequence.

This separation between event timing and shot structure is more clearly illustrated by the speech-across-shots example. The cyclist's utterance remains continuous while the scene transitions from a medium close view to a wide shot, showing that semantic events can extend across shot boundaries without being interrupted by visual transitions. The narrated sequence further combines speech with successive actions across multiple shots, while preserving the same character and scene context.

The same control extends beyond speech. In the final example, different camera behaviours are assigned to consecutive shots, with a camera roll followed by a zoom while the scene remains consistent. Together, these examples show that SceneGen can represent speech, actions, shot transitions, and camera motion at their appropriate temporal scales, which is central to fine-grained multi-shot generation.

3.4 Ablation Study

Table 3 examines the contributions of shot-level and event-level temporal modeling. The results show a clear division of roles: Shot-Aware RoPE mainly improves shot organization, while Event-Aware Attention provides stronger control over temporally localized semantic events. Their combination enables the model to coordinate both temporal scales.

Shot-Aware RoPE. Adding Shot-Aware RoPE reduces CTE from 4.27 to 1.52 and improves CC from 3.79 to 4.17, while STE changes only slightly from 3.62 to 3.26. This pattern indicates that explicit shot positions mainly benefit shot transitions and cross-shot consistency, but provide limited guidance for fine-grained event timing.

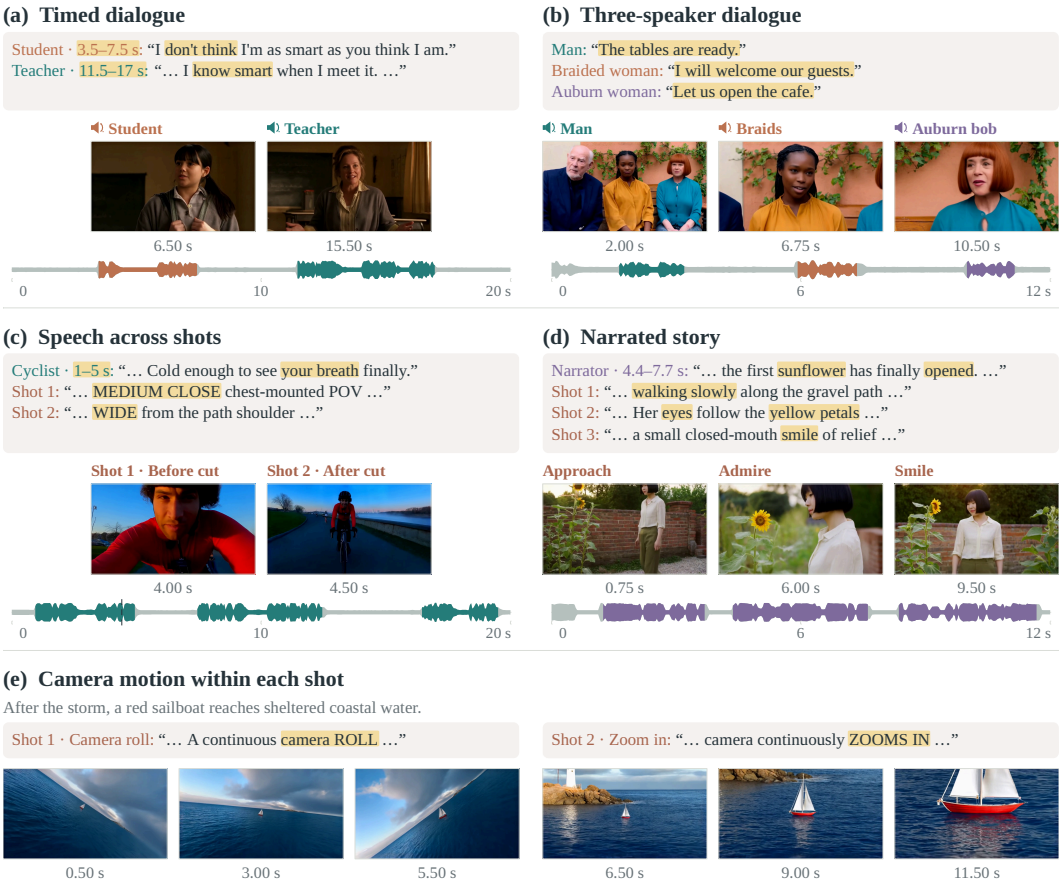


Figure 4: **Qualitative examples of fine-grained temporal control over speech, identities, actions, and camera motion.** Amber highlights requested instructions. Waveform colours indicate different speakers, while grey denotes other audio.

Table 3: **Ablation study of temporal conditioning and event encoding.** All variants are initialised from LTX-2.3 and use the same training budget.

Configuration	WER ↓	STE ↓	Desync ↓	CTE ↓	PF ↑	CC ↑
SFT only	1.79	3.62	0.43	4.27	3.96	3.79
SFT + Shot-Aware RoPE	2.85	3.26	0.39	**1.52**	3.91	4.17
SFT + Event-Aware Attn.	**0.68**	0.36	0.31	3.67	**4.26**	4.21
SceneGen	0.79	**0.23**	**0.28**	1.97	4.19	**4.31**
Joint event encoding	2.48	3.73	0.38	3.95	3.79	3.89

Event-Aware Attention. Event-Aware Attention shows the complementary effect. It reduces STE from 3.62 to 0.36, WER from 1.79 to 0.68, and Desync from 0.43 to 0.31, while improving PF from 3.96 to 4.26. In contrast, its CTE remains relatively high at 3.67. These results show that temporal routing is particularly effective for controlling event content and timing, but is less effective at organizing the global shot structure.

Combining Shot and Event Conditioning. The full model combines both components and achieves the best STE of 0.23, Desync of 0.28, and CC of 4.31, while maintaining low CTE and strong PF. Although individual variants perform slightly better on a few metrics, the full model provides the strongest overall balance across event-level and shot-level control. This supports the complementary roles of Shot-Aware RoPE for shot structure and Event-Aware Attention for temporally localized semantics.

Table 4: **Human evaluation of multi-shot visual control and audio generation.** Twenty-five participants rate anonymised clips from 1 to 5. The left table evaluates multi-shot visual generation, while the right table evaluates ambient sound generation and audio instruction following. Higher values indicate better performance.

Model	PF ↑	Aesthetics ↑	CC ↑
HoloCine	<u>3.96</u>	**4.32**	3.84
MultiShotMaster	3.84	3.96	<u>4.16</u>
SceneGen	**4.12**	<u>4.28</u>	**4.40**

Model	Env. ↑	Audio Quality ↑	PF ↑
LTX-2.3	<u>3.28</u>	**4.16**	<u>3.56</u>
Ovi	3.20	3.96	3.16
SceneGen	**3.64**	<u>4.04</u>	**3.72**

Independent Event Encoding. Replacing independent event encoding with joint encoding substantially degrades performance, increasing STE from 0.23 to 3.73 and WER from 0.79 to 2.48, while also reducing PF and CC. This suggests that mixing different event descriptions before temporal routing weakens their temporal correspondence. Independent encoding therefore helps preserve event-specific representations before they are assigned to their respective temporal intervals.

3.5 Human Evaluation

Human evaluation further supports the quantitative results. For multi-shot visual generation, SceneGen achieves the highest PF and CC scores while maintaining competitive visual aesthetics, demonstrating improved visual control and cross-shot character consistency.

For audio generation, we evaluate ambient sound with human ratings because its scene-level appropriateness is difficult to measure using automatic metrics. SceneGen achieves the highest Env. and PF scores while maintaining comparable audio quality, indicating stronger audio controllability without compromising perceptual quality.

4 Related Work

Multi-shot Video Generation. Recent multi-shot video models introduce explicit shot-level control through localized conditioning. ShotAdapter uses transition tokens and local attention masks for shot-specific prompts (Kara et al., 2025), while HoloCine applies window cross-attention with sparse inter-shot attention (Meng et al., 2025). MultiShotMaster further introduces rotary phase shifts at shot transitions for structured multi-shot generation (Wang et al., 2025b). These methods establish explicit temporal structure as an important component of controllable multi-shot video generation.

Multi-shot Audio-Video Generation. Recent work further considers audiovisual continuity across multiple shots. UnityShots maintains visual and vocal identity across shots (Huang et al., 2026), while MAVIN uses hierarchical descriptions and boundary-aware routing for temporally structured audiovisual control (Liu et al., 2026). CineDance extends this setting to long-form multi-shot audiovisual generation with structured multimodal annotations (Chen et al., 2026). Detailed Survey is listed in Appendix A.

5 Conclusion

We presented SceneGen for fine-grained multi-shot audio-video generation with independently timed visual and audio control within a unified temporal modeling framework. By combining Shot-Aware RoPE with Event-Aware Attention, SceneGen separates shot-level structure from temporally localized events while preserving their coordination within a shared sequence. Trained on SceneGen-Data, the model improves speech timing, instruction following, and character consistency on SceneGen-Bench while maintaining audiovisual quality. Ablation studies and human evaluation further support the effectiveness of this design. These results highlight the importance of matching conditioning granularity to the temporal structure of multi-shot audio-video generation.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript for clarity and readability and to assist in writing LaTeX code for figures and tables. We also used AI deep research functionality to support a broad search for relevant literature for the related work section. We reviewed all AI-assisted work, including the edited text and generated LaTeX code. For the literature search, we manually verified the papers identified through AI-generated lists before citing them in the manuscript. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, data construction, and experimental protocol to facilitate reproducibility. The architecture and temporal conditioning mechanisms of SceneGen, including Shot-Aware RoPE and Event-Aware Attention, are specified in Sec. 2, with their formulations and implementation details given in Secs. 2.2 and 2.3. The construction and annotation procedure of SceneGen-Data is described in Sec. 2.1, with additional dataset statistics, filtering criteria, and annotation examples provided in Appendix B. Training, optimisation, and inference configurations are documented in Sec. 3.1 and Appendix C, while baseline configurations and evaluation definitions are provided in Appendix C.2 and Appendix C.3. The composition and temporal-control coverage of SceneGen-Bench, together with a representative benchmark record and additional qualitative results, are provided in Appendix D and Appendix E. We will release the dataset and trained models to support reproduction and further research.

REFERENCES

Alibaba Cloud. Wan3.0 video generation, 2026. URL https://docs.modelstudio.console.alibabacloud.com/en/model-studio/wan3-video-generation-guide.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *24th Annual Conference of the International Speech Communication Association*, pp. 4489–4493, 2023.

ByteDance Seed. One-take creation, flexible referencing: Introducing Seedance 2.5, 2026. URL https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 13320–13331, 2024.

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*, 2026.

Shihao Cheng, Jiaxu Zhang, Quanyue Song, Shansong Liu, Zhizhi Guo, Xiao-Lei Zhang, Chi Zhang, Xuelong Li, and Zhigang Tu. Unison: Harmonizing motion, speech, and sound for human-centric audio-video generation. In *Computer Vision – ECCV*, pp. 96–115, 2026.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. FineVideo, 2024. URL https://huggingface.co/datasets/HuggingFaceFV/finevideo.

Haokun Gui, Senqiao Yang, Mingkang Zhu, Meng Chu, Sitong Wu, Changsheng Lu, Zihao Wang, Zhuotao Tian, and Jiaya Jia. Visionleaf: Entropy-guided leaf-first reasoning for efficient and accurate think-with-image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5187–5198, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy

Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Díez, and Lukás Burget. Leveraging self-supervised learning for speaker diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 1–5, 2025a.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *International Conference on Learning Representations, ICLR*, 2025b.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR*, 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.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision - ECCV*, volume 12349, pp. 709–727, 2020.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 5325–5329, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 28405–28415, 2025.

LAION-AI. LAION-Aesthetics Predictor V1, 2022. URL `https://github.com/LAION-AI/aesthetic-predictor`.

Yulin Li, Haokun Gui, Ziyang Fan, Junjie Wang, Bin Kang, Bin Chen, and Zhuotao Tian. Less is more, but where? dynamic token compression via llm-guided keyframe prior. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems*, 2025.

Lightricks. LTX-2.3 model card, 2026. URL `https://huggingface.co/Lightricks/LTX-2.3`.

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.

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*, 2026.

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, 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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139, pp. 8748–8763, 2021.

RapidAI Team. Rapid OCR: Ocr toolbox, 2021. URL https://github.com/RapidAI/RapidOCR.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 10219–10228, 2023.

SII-OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu, Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Quanyue Song, Yishan He, Yanfei Zhang, Shihao Cheng, Zhixiang He, Zhizhi Guo, Chi Zhang, Xuelong Li, and Caigui Jiang. Interactiveavatar: Real-time streaming video generation for consistent and intent-aware avatars, 2026.

Tomás Soucek and Jakub Lokoc. Transnet V2: an effective deep network architecture for fast shot transition detection. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (eds.), *Proceedings of the 32nd ACM International Conference on Multimedia, MM*, pp. 11218–11221, 2024.

Zachary Teed and Jia Deng. RAFT: recurrent all-pairs field transforms for optical flow (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pp. 4839–4843, 2021.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *arXiv preprint arXiv:2509.25131*, 2025a.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025b.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 8428–8437, 2025c.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 20087–20097, 2023.

Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. *arXiv preprint arXiv:2508.11484*, 2025.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. LVD-2M: A long-take video dataset with temporally dense captions. In *Advances in Neural Information Processing Systems*, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, 2023.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, abs/2308.08089, 2023.

Yanhong Zeng, Jianlong Fu, and Hongyang Chao. Learning joint spatial-temporal transformations for video inpainting. In *Computer Vision - ECCV*, volume 12361, pp. 528–543, 2020.

Jun Zhan, Chen Yang, Yitian Gong, Donghua Yu, Kuangwei Chen, Wenbo Zhang, Kexin Huang, Qi Luo, Zhe Xu, Ying Zhu, Jin Wang, Tengyue Zhang, Qi Chen, Cheng Chang, Songlin Wang, Junqi Dai, Jiasheng Ye, Xiaogui Yang, Tianyi Liang, Xiangyu Peng, Zhaoye Fei, Shimin Li, Qinyuan Cheng, Xie Chen, Xinchi Chen, and Xipeng Qiu. Omnivae: An audio-video VAE with cross-modal alignment for joint generation. *arXiv preprint arXiv:2607.23855*, 2026.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Jiahe Zhang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, and Xiu Li. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. *arXiv preprint arXiv:2507.09862*, 2025.

A Additional Related Work

Controllable Video Generation. Controllable video generation extends text-to-video models with additional conditions for camera motion, spatial trajectories, scene structure, and recurring entities. CameraCtrl introduces explicit camera trajectory conditioning to control viewpoint changes during video generation (He et al., 2025). DragNUWA combines text, image, and trajectory conditions, allowing motion paths to guide object and scene dynamics (Yin et al., 2023). Beyond local motion control, VideoDirectorGPT decomposes video generation into scene planning and generation, using structured descriptions of scenes and recurring entities to support longer visual narratives (Lin et al., 2023). These studies broaden video conditioning from global semantic prompts to explicit spatial, motion, and narrative structures.

Joint Audio-Video Generation. Joint audio-video generation aims to synthesize visual and acoustic content within a shared generative process. MM-Diffusion couples modality-specific audio and video denoisers through cross-modal attention, enabling information exchange during diffusion (Ruan et al., 2023). Ovi uses bidirectional cross-modal interactions between audio and video streams to jointly model the two modalities (Low et al., 2025). LTX-2 develops a unified audiovisual generation architecture in which audio and visual representations are jointly processed during generation (HaCohen et al., 2026). OmniVAE instead studies audiovisual alignment in a shared latent representation and supports generation from the resulting multimodal latent space (Zhan et al., 2026). Collectively, these approaches (Cheng et al., 2026; Song et al., 2026) explore architectural mechanisms for generating audio and video.

Temporally Structured Video Datasets. Video datasets differ substantially in the temporal granularity of their annotations. Panda-70M provides large-scale video-text pairs with clip-level descriptions and broad visual coverage (Chen et al., 2024). LVD-2M introduces dense temporal captions for long videos, describing content changes across multiple temporal segments rather than representing the complete video with a single caption (Xiong et al., 2024). Cine250K focuses on cinematic video structure and provides shot boundaries together with shot-level descriptions, supporting the study of transitions and composition across consecutive shots (Wu et al., 2025). These datasets (Han et al., 2025b; Huang et al., 2020; Yang et al., 2023) progressively increase annotation granularity from whole clips to temporally localized segments and individual shots.

Speech and Speaker Annotations in Video. Audiovisual datasets additionally provide supervision for speech and speaker correspondence. SpeakerVid-5M associates visible characters with speakers and includes transcription and speaker-turn information (Zhang et al., 2025). Such annotations connect spoken content with character identity and temporal speech structure using specific models (Wang et al., 2025a; Gui et al., 2026; Li et al., 2025), supporting tasks that require both visual person understanding and speech modeling. More broadly, speaker-aware annotations complement visual scene descriptions by representing who speaks and when speech occurs within a video.

B Training Data and Annotations

The Sec. 2.1 introduces SceneGen-Data as a multi-shot audio-video dataset designed to support two complementary temporal structures. Visual content is organized by shot boundaries, while speech and sound events retain independent intervals on the continuous scene timeline. Appendix B provides additional details on the data distribution, filtering process, and annotation format used to support this design.

SceneGen-Data contains approximately 194K multi-shot training scenes collected from Panda-70M, Koala-36M, and internally collected Internet videos. The collection process starts from 6,945 deduplicated source videos and 339,823 candidate scenes. Each retained example preserves consecutive source footage together with its original aligned audio. The resulting content covers television and film dialogue, vlogs and tutorials, public speaking, music and performance, sports, cooking, and natural scenery. Figure 5 illustrates representative source content.

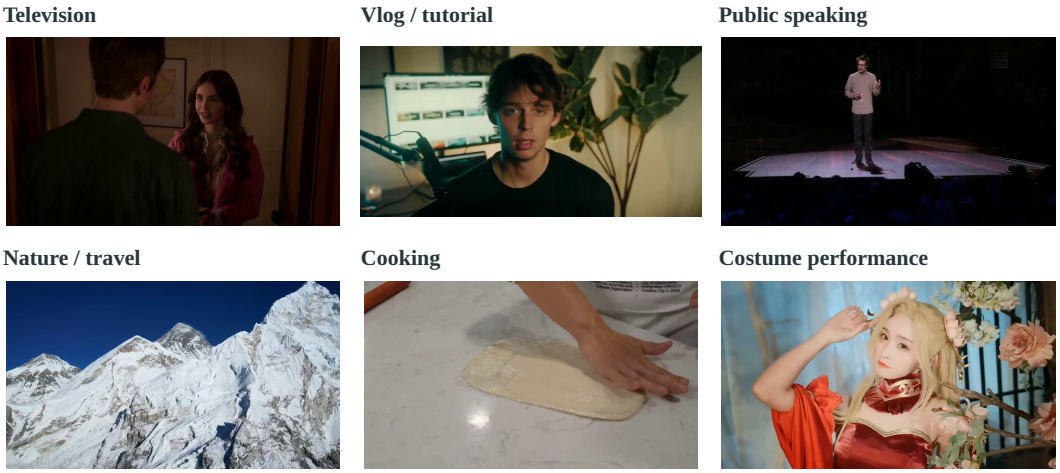


Figure 5: **Examples of training-video content.** Original source frames illustrate television, vlogs and tutorials, public speaking, nature and travel, cooking, and costume performance.

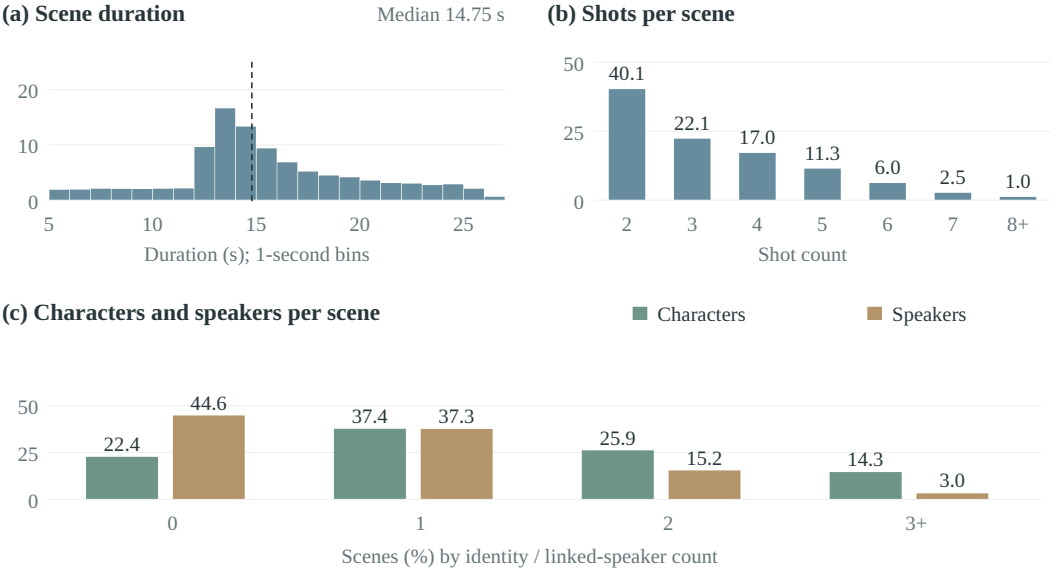


Figure 6: **Duration, shot, and character distributions of the screened data pool.** All panels are computed from the same records before the final training selection. The duration histogram uses one-second bins and marks the median. The remaining panels summarize shot counts, linked character identities, and character-linked speakers. Values are reported as percentages of scenes.

B.1 Scene, Shot, and Character Statistics

Before the final training selection, 194,239 multi-shot records pass the available content, cache, and duration checks. This screened pool contains 646,326 shots and 838.30 hours of source footage. The mean scene duration is 15.54 seconds and the median is 14.75 seconds. Among these scenes, 73.1% last between 10 and 20 seconds, while 79.2% contain two to four shots. These statistics are consistent with the intended setting of short but structurally nontrivial multi-shot sequences.

The screened pool contains diverse character configurations. Among all scenes, 22.4% contain no annotated character identity, 37.4% contain one identity, and 25.9% contain two identities. Character counts are defined at the scene level and may therefore exceed the number of characters visible within an individual shot.

Speaker annotations describe speech associated with linked character identities. In the screened pool, 37.3% of scenes contain one character-linked speaker, while 18.1% contain at least two. Scenes with zero linked speakers may still contain narration, off-screen speech, or other utterances

Table 5: **Filtering criteria and rejection rates.** Percentages indicate the fraction of multi-shot videos rejected by each criterion. Filtering criteria are evaluated independently and are not mutually exclusive; therefore, a video may be rejected by more than one criterion and the reported percentages do not sum to the overall rejection rate.

Criterion	Rejected Rate(%)
Visual quality	∼15.0
Audio quality	∼11.0
Motion	∼4.0
Black borders	∼0.6
Text overlays	∼0.5
Black frames	∼0.7
Crowded shots	∼15.0

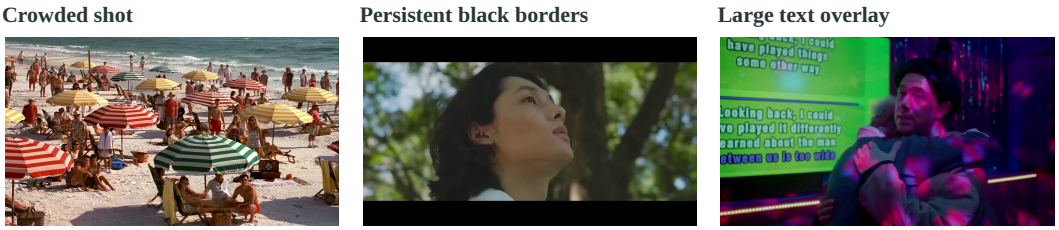


Figure 7: **Examples rejected during content screening.** Representative source frames show crowded compositions, persistent black borders, and large text overlays. The displayed frames preserve the original visual content before filtering.

that are not assigned to a visible character identity. This distinction keeps visual identity association and speech timing as separate annotation dimensions, allowing each to reflect the information explicitly represented in the source content.

B.2 Filtering Criteria and Rejected Scenes

The filtering stage removes scenes that are unsuitable for fine-grained audiovisual supervision. As described in Sec. 2.1, DOVER++ measures visual quality, Audiobox Aesthetics measures audio quality, and RAFT is used to identify scenes with insufficient motion. Additional content checks identify persistent black borders, large text overlays, black frames, and overly crowded compositions.

The main collection rejects around 23% candidate scenes through the combined quality, motion, and border checks. Table 5 reports the recorded content-screening criteria separately. A scene may fail more than one test, so the individual rejection counts should not be summed. Under the recorded content-screening configuration, 214,022 multi-shot videos are evaluated and 16,824 distinct videos are rejected by at least one content criterion.

The filtering stages serve different purposes. Quality and motion checks remove globally unsuitable media, while content screening targets structural artifacts that can interfere with scene annotation or generation supervision. Figure 7 shows representative examples rejected by the latter stage.

B.3 Annotation Structure

After filtering and character association, each scene is annotated at three complementary levels. A global description summarizes scene-level appearance and context. Shot captions describe local viewpoint, camera behavior, visible characters, and actions. Speech and sound records retain their own temporal intervals on the original scene timeline.

This representation is important because camera transitions and audio events do not necessarily share the same boundaries. A shot caption changes when the camera cuts, while an utterance or

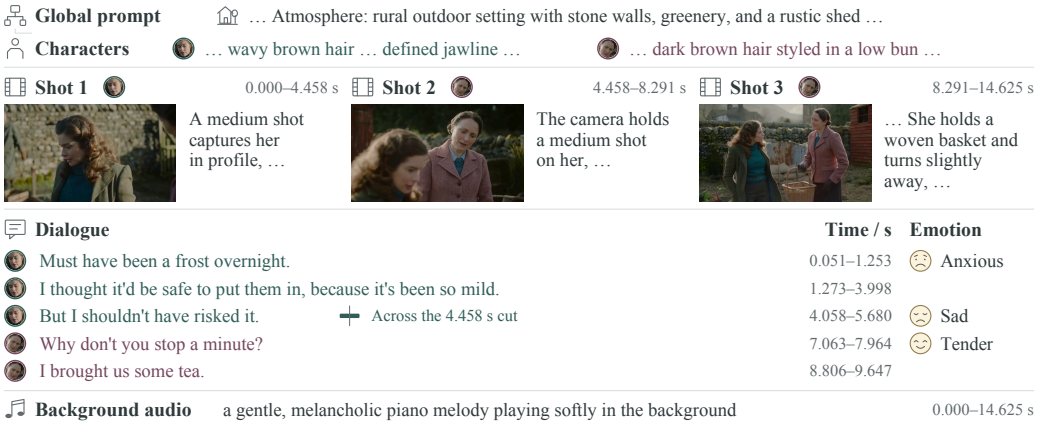


Figure 8: **Shot-level visual annotations and independently timed audio events.** Original source frames are paired with shot captions. Character labels identify recurring speakers across views. Speech may remain active across a camera cut, while speech and sound preserve their own temporal intervals on the original scene timeline.

sound event can begin within one shot and continue into another. The annotation format therefore provides the supervision required by Shot-Aware RoPE and Event-Aware Attention without forcing the two temporal structures to share a common segmentation.

Figure 8 provides a concrete annotation example. Original frames are paired with shot-level visual descriptions, while speech and sound remain independently timed. Character labels identify recurring speakers across views. The example illustrates that an utterance can cross a camera cut while retaining one continuous event interval.

This annotation structure directly matches the conditioning strategy used by SceneGen. Shot boundaries describe changes in visual composition, while event intervals determine when speech and sound descriptions become accessible to the model. The dataset therefore preserves the temporal distinction required for fine-grained control within and across shots.

C Training and Experimental Setup

C.1 Optimisation and Length Budgets

We initialise the 22B audiovisual denoiser from LTX-2.3 and fine-tune it on the approximately 194K-scene training set. Video latents, audio latents, and text features are cached before optimisation. The corresponding encoders and decoders remain fixed during training. Table 6 summarises the configuration used for the experiments reported in Sec. 3.1.

Length-budget sampling. Each example has a fixed cached temporal length. Examples are sorted by temporal length and organized into groups of 128 for individual optimiser updates. The selected length budget is shared across devices and accumulation microsteps within the same update.

Different groups use the same resolution, optimiser, and learning-rate schedule. Only the temporal budget varies between updates. This strategy limits unnecessary computation caused by placing unusually long and short scenes in the same update. The grouping procedure depends only on cached scene length and does not require unrelated scenes to share shot boundaries or event times.

Joint and conditional generation tasks. Global text, shot captions, and audio-event descriptions are encoded separately. Their accessibility is determined by the temporal routing mechanism introduced in Sec. 2.3. This preserves the distinction between scene-level context, shot-local visual instructions, and independently timed audio events throughout training.

Inference settings. The benchmark outputs reported in the paper use 30 denoising steps at 1280×704 and 24 fps. Requested shot intervals and audio-event intervals are supplied directly to SceneGen.

Table 6: **Training configuration for SceneGen on SceneGen-Data.** The table reports the optimisation setup, task-sampling probabilities, conditioning dropout, and temporal-control parameters used during training.

Setting	Value
Initialisation	LTX-2.3 22B
Trainable component	Audiovisual denoiser
Training data	Approximately 194K multi-shot scenes
Resolution	1280×704
Frame rate	24 fps
Hardware	64 NVIDIA H20 GPUs
Global batch size	128
Optimiser	AdamW
Weight decay	0.01
Peak learning rate	10^{-5}
Learning-rate schedule	500-step warm-up followed by cosine decay
Gradient clipping	Global norm 1.0
Precision	BFloat16
Memory optimisation	ZeRO-3
Global condition dropout	0.1
Shot condition dropout	0.1
Audio-event condition dropout	0.1
Shot positional shift	0.5 s
Event onset tolerance	0.02 s

Video generation uses text guidance 3 and modality guidance 3. Audio generation uses text guidance 7 and modality guidance 3.

C.2 Baseline Models

The baselines are evaluated using their publicly available models and native inference procedures. When a model has a fixed native duration, benchmark timing is mapped to that duration for evaluation. Video-only methods are evaluated only on applicable visual and multi-shot metrics.

MultiShotMaster. MultiShotMaster-14B (Wang et al., 2025b) extends a video generator with Multi-Shot Narrative RoPE for shot transitions and position-aware reference conditioning. It accepts a global description, separate shot captions, and explicit shot groups. The recorded workflow uses 50 sampling steps, CFG 5, 1280×720 resolution, and 15 fps.

HoloCine. HoloCine-14B-full (Meng et al., 2025) generates a multi-shot sequence from global scene context and local shot descriptions. We serialize the visual script using its native prompt format and provide the requested shot-cut frames. The recorded workflow uses 50 sampling steps, CFG 5, 832×480 resolution, and 15 fps.

LTX-2.3. LTX-2.3 (Lightricks, 2026) is the 22B joint audiovisual backbone used to initialise SceneGen. The baseline receives scene text through its native conditioning interface and uses its HQ generation workflow.

Ovi. Ovi 1.1 (Low et al., 2025) couples video and audio diffusion transformers through bidirectional cross-modal fusion. Visual content, dialogue, and sound requirements are serialized into its native text prompt. Its recorded workflow uses 1280×704 resolution at 24 fps, 50 UniPC steps, video guidance 4, and audio guidance 3.

OmniVAE T2AV. OmniVAE (Zhan et al., 2026) aligns audio and video representations within a joint variational autoencoder for audiovisual generation. We evaluate its text-to-audio-video checkpoint using 50 sampling steps, and CFG 4.

Table 7: **Evaluation metrics used in the main experiments.** The metrics separately evaluate generation quality, speech content and timing, audiovisual synchronization, and multi-shot control.

Metric	Definition
TA	Text-video alignment measured with ViCLIP (Wang et al., 2024).
AES-V	Visual aesthetics measured with the LAION aesthetic predictor (LAION-AI, 2022; Radford et al., 2021).
CU	Audio content usefulness measured with Audiobox Aesthetics (Tjandra et al., 2025).
PQ	Audio production quality measured with Audiobox Aesthetics (Tjandra et al., 2025).
WER	Word error rate computed from Whisper large-v3 transcripts against the requested utterance content.
Incorrect	Percentage of generated utterances whose recognized speech does not match the requested utterance.
Desync	Audio-visual speech synchronization offset in seconds measured with Synchformer (Iashin et al., 2024).
PF	Five-point GPT-based score for following the requested visual and temporal instructions.
CC	Five-point GPT-based score for character consistency across shots.
CTE	Shot-boundary timing error in frames using TransNetV2 (Soucek & Lokoc, 2024).
STE	Normalized start and end boundary error for requested utterances.

MOVA. MOVA-720p (SII-OpenMOSS Team et al., 2026) is evaluated using the authors' constant-white-frame T2VA workflow.

C.3 Evaluation Definitions and Reporting

The main experiments evaluate three complementary aspects: audiovisual quality, speech generation, and fine-grained multi-shot control. Table 7 summarises the metrics used in Table 2. Higher values are preferred for TA, AES-V, CU, PQ, PF, and CC. Lower values are preferred for WER, Incorrect, Desync, CTE, and STE.

Shot timing and audiovisual synchronization. CTE measures the temporal displacement between requested and detected shot boundaries. Generated cuts are obtained with TransNetV2 and compared with the target boundaries in frame units. Desync measures the temporal offset between visible speech activity and generated speech audio using Synchformer. These metrics evaluate different structures: CTE measures camera-transition timing, while Desync measures local audiovisual synchronization.

Prompt following and character consistency. PF evaluates whether the generated sequence follows the requested visual content, actions, camera behavior, and temporal instructions. CC evaluates whether recurring characters preserve their visual identity across changes in shot, scale, and viewpoint. Both metrics use a five-point evaluation scale.

D SceneGen-Bench

D.1 Content and Control Coverage

SceneGen-Bench contains 100 fixed multi-shot prompts designed to evaluate the temporal structures targeted by SceneGen. The prompts are generated with GPT and subsequently manually reviewed and curated to define the evaluation categories and balance the coverage of different test types. They are constructed specifically for benchmarking rather than sampled or retrieved from the training data. Each prompt provides a global scene description, an ordered sequence of visual shots, and independently timed speech and sound events. Across the benchmark, the prompts specify 302 shots, 202 camera cuts, 160 utterances, and 119 sound events.

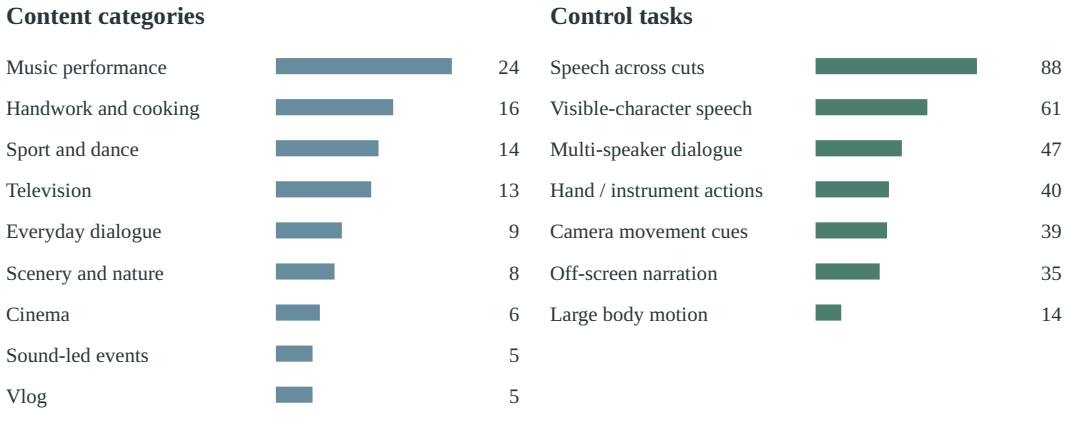


Figure 9: **Content and control coverage of SceneGen-Bench.** Ranked bars summarize primary content categories and selected control tasks. Each prompt has one primary content category and may belong to multiple control subsets. Values denote prompt counts within the shared 100-prompt benchmark.

The benchmark is constructed to test whether a model can follow local shot instructions while preserving speech and sound events whose intervals do not coincide with shot boundaries. Figure 9 summarizes the main content categories and control requirements.

Music performance contributes 24 prompts, handwork and cooking contribute 16, sports and dance contribute 14, television contributes 13, everyday dialogue contributes nine, scenery and nature contribute eight, cinema contributes six, and vlogs and sound-led scenes contribute five each. Together, these categories cover performance, object manipulation, large body motion, conversation, environmental change, and multi-view narrative structure.

D.2 Temporal Structure and Participants

Requested durations are 15, 18, 20, and 22 seconds, with a mean duration of 19.85 seconds. Ninety-eight prompts contain three shots and two contain four shots. Eight prompts contain no character identity, 41 contain one, 50 contain two, and one contains three.

Character-linked speaker counts follow a different distribution. Thirty-nine prompts contain no linked speaker, 14 contain one, 46 contain two, and one contains three. The distinction between characters and speakers is intentional because a visible character does not necessarily speak, while narration may occur without a visible speaker.

D.3 Cross-Shot Audiovisual Events

The benchmark places particular emphasis on events whose timing is independent of camera transitions. Sixty-one prompts request speech from a visible character, while 47 contain multiple speakers. Thirty-five contain off-screen narration.

Eighty-eight prompts contain speech that crosses at least one camera cut, and 101 of the 160 requested utterances cross a cut. Every benchmark prompt contains at least one sound event that spans a camera transition, while 95 contain overlapping speech and sound intervals. These examples directly evaluate whether a model can preserve event timing when the active visual shot changes.

D.4 Motion and Fine-Grained Actions

The benchmark also evaluates visual actions at different spatial and temporal scales. Fourteen sports and dance prompts emphasize large body motion such as skating, stepping, and turning. Music performance and handwork together contribute 40 prompts involving instrument playing, tool use, cooking, or object manipulation.

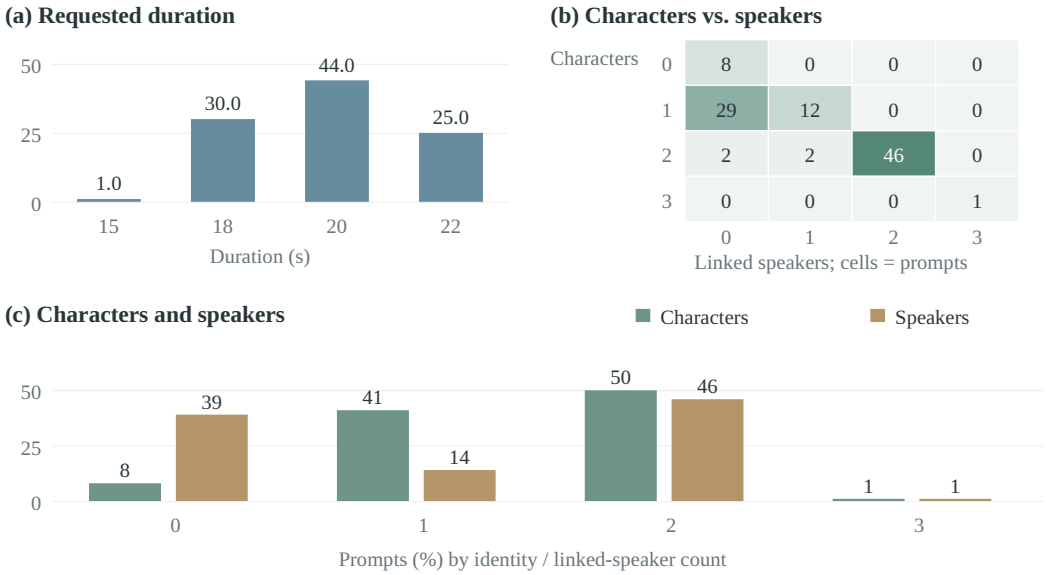


Figure 10: **Temporal structure and participants in SceneGen-Bench.** Duration bars summarize the 100 benchmark prompts. Additional panels compare character and linked-speaker counts and show their joint distribution. Counts are numerically equivalent to percentages because the benchmark contains 100 prompts.

Thirty-nine prompts explicitly specify camera movement. PF measures whether the requested action and visual behavior are followed, CTE measures the timing of shot transitions, and CC measures whether recurring characters remain visually consistent across different views. These metrics therefore evaluate complementary aspects of multi-shot control rather than reducing performance to a single score.

D.5 Combined Conditioning Tasks

Most benchmark prompts combine camera transitions, character actions, speech, and environmental sound within the same sequence. This structure is intended to test whether the model can coordinate multiple forms of control without assigning all instructions to the same temporal scope.

Two prompts explicitly describe audio-driven motion, while three describe video-driven sound events. These labels describe the intended conditioning relationship within the benchmark and are separate from the joint, video-to-audio, and audio-to-video task probabilities used during training.

D.6 A Benchmark Record

Figure 11 shows a concrete example from SceneGen-Bench. The example contains three temporally ordered shots together with independently specified audio events. The winding sound is localized to the first shot, while the music-box melody begins in the second shot and continues across the cut at 14.0 seconds into the third shot. This example illustrates how SceneGen-Bench represents shot structure and event timing separately, allowing audio events to begin, end, or persist independently of shot boundaries.

E Additional Cross-Shot Generation Results

Figures 12 to 14 provide additional qualitative examples of multi-shot generation. Each row shows two temporally ordered frames from the same sequence, highlighting changes in viewpoint, framing, appearance, or action while remaining consistent with the corresponding visual instruction.

These examples complement the main qualitative results by illustrating shot-specific instruction following and cross-shot character and scene consistency. Representative frames are shown as visual evidence and are not intended to replace the quantitative evaluation.

```
{
  "id": "audio_driven_music_box_char_3shot_20s",
  "duration_sec": 20.0,
  "global_caption": "[character1]: female, adult, medium build, with a calm and
  composed appearance.\nAtmosphere: A warmly lit, quiet study with an intimate and
  contemplative mood, centered around a small mechanical music box.",
  "language": "en",
  "shot_count": 3,
  "shot_annotations": [
    {
      "shot_id": "shot_0000",
      "start_sec": 0.0,
      "end_sec": 5.5,
      "visible_characters": [
        "character1"
      ],
      "shot_caption": "A medium view shows [character1] holding a small music box in
      a warmly lit study. She begins winding the key at 2.5 seconds and continues
      until 5.0 seconds, while the surrounding room remains quiet and still."
    },
    {
      "shot_id": "shot_0001",
      "start_sec": 5.5,
      "end_sec": 14.0,
      "visible_characters": [
        "character1"
      ],
      "shot_caption": "An over-the-shoulder view focuses on the music box in front
      of [character1]. At 6.5 seconds, the internal mechanism begins to move and the
      music box starts playing, drawing attention to its delicate mechanical
      details."
    },
    {
      "shot_id": "shot_0002",
      "start_sec": 14.0,
      "end_sec": 20.0,
      "visible_characters": [
        "character1"
      ],
      "shot_caption": "The camera returns to a medium view of [character1] as she
      leans back slightly and listens to the music box. The delicate melody
      continues through the end of the shot, maintaining the warm and quiet
      atmosphere of the study."
    }
  ],
  "voice_annotation": [
    {
      "type": "sfx",
      "content": "Quiet study ambience with faint clock ticking in the background.",
      "start_sec": 0.0,
      "end_sec": 20.0
    },
    {
      "type": "sfx",
      "content": "Mechanical clicking sounds as the music-box key is wound.",
      "start_sec": 2.5,
      "end_sec": 5.0
    },
    {
      "type": "sfx",
      "content": "A delicate, continuous music-box melody.",
      "start_sec": 6.5,
      "end_sec": 20.0
    }
  ]
}
```


Figure 11: **Example benchmark annotation.** An example of the structured multi-shot annotation format used in SceneGen-Bench, including global scene information, temporally localized shot descriptions, visible character identities, and audio events.

People and changing viewpoints

(a) Reading a letter by the window



Shot 1 · 3.50 s

… a TIGHT medium close-up on [character1]'s face from front-on …



Shot 2 · 7.25 s

… looking toward the window … at the reading table with the window filling the right side of the frame …

(b) A conversation in a diner



Shot 1 · 2.50 s

… a locked MEDIUM two-shot across the booth … [character1] on the left and [character2] on the right …



Shot 2 · 14.00 s

… a MEDIUM reverse on [character2] … wearing the same burgundy cable-knit sweater …

(c) Viewing a painting



Shot 1 · 2.50 s

… view from behind the woman … facing the red painting …



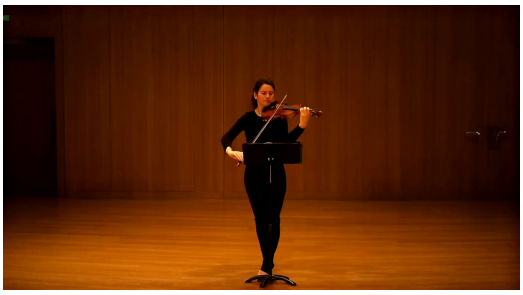
Shot 2 · 4.25 s

… a close side view of the woman's face as she looks at the painting …

Figure 12: **Additional cross-shot examples with changing viewpoints.** The examples show a letter-reading scene, a diner conversation, and an art-gallery visit. Each row presents two temporally ordered views from the same generated sequence and highlights the corresponding visual instruction.

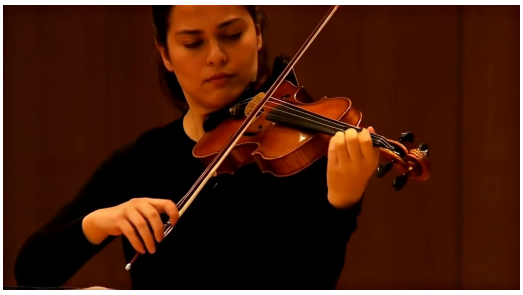
Performance and visual detail

(a) A solo violin performance



Shot 1 · 2.50 s

… a locked WIDE of the stage … the violin at her shoulder …



Shot 2 · 9.50 s

… a MEDIUM three-quarter on [character1] … the bow drawing a long sustained stroke across the strings …

(b) A dancer at the barre



Shot 1 · 2.00 s

… on [character1] at the barre … warm sun banding across the sprung wood floor …



Shot 2 · 16.25 s

… a MEDIUM close-up on [character1] …

(c) A cat beside the window



Shot 1 · 1.50 s

… A wide side view … ginger cat resting on the blue cushion. …



Shot 2 · 4.50 s

… a close view of the ginger cat's face and whiskers …

Figure 13: **Additional cross-shot examples with performance and local visual detail.** A violinist, a dancer, and a cat illustrate transitions between wider and closer views while preserving the subject and scene context across shots.

Landscapes and camera angles

(a) From the bamboo path to the canopy



Shot 1 · 2.00 s

… a WIDE static shot down the footpath … the path receding into the green depth …



Shot 2 · 15.00 s

… a WIDE static shot straight up the bamboo trunks … tall green stalks converging toward a bright canopy …

(b) A hiker among the dunes



Shot 1 · 10.00 s

… a TIGHT medium close-up on [character1]'s face … his hat brim shading his eyes …



Shot 2 · 16.00 s

… a WIDE overhead drone-like shot high above the dune ridge … [character1] as a tiny figure …

(c) Roses in a sunlit garden



Shot 1 · 1.50 s

… A wide view shows the flowering rose bushes in the garden. …



Shot 2 · 4.50 s

… a close view of a pink rose with green leaves behind it …

Figure 14: **Additional cross-shot examples with landscapes and camera-angle changes.** The examples include a bamboo path and canopy view, a hiker shown from different viewpoints, and a transition from a rose garden to local flower detail.