SceneShifter: Training-free Multi-Scene Temporal Control for Audio-driven Human Animation

Jianzhi Long[1] **Rong-Cheng Tu**[2*] **Hao Guan**[2] **Siyuan Liang**[2]
Shunyu Liu[2] **Xiao Luo**[3] **Dacheng Tao**[2*]

[1]The University of Sydney, New South Wales, Australia
[2]Generative AI Lab, College of Computing and Data Science,
Nanyang Technological University, Singapore, Singapore
[3]University of Wisconsin–Madison, Wisconsin, United States

Abstract

Although audio-driven human animation has achieved impressive realism, it still lacks effective **Multi-Scene Temporal Control**: orchestrating background transitions at precise timestamps while preserving a coherent foreground subject. We show that this challenge stems from the diffusion transformer's self-attention mechanism, and addressing it requires two capabilities: (1) **Semantic-Aware Temporal Decoupling**, which isolates background attention across scenes while retaining cross-scene foreground attention for subject consistency; and (2) **Foreground Motion Localization**, which accurately tracks the foreground subject across latent frames, especially under large motion. To address these challenges, we introduce SceneShifter, a training-free framework for precise multi-scene control in audio-driven human animation. SceneShifter guides spatio-temporal self-attention by suppressing cross-scene attention among background tokens while preserving cross-scene attention among foreground tokens. To support this guidance under large motion, SceneShifter extracts dynamic foreground masks from the self-attention layers and heads that best capture subject trajectories. We further introduce **SceneShifterBench**, a benchmark designed to evaluate scene-timing accuracy, foreground preservation, and visual quality under large motion, multiple subjects. Experiments show that SceneShifter achieves frame-level scene timing, strong subject preservation, and high visual quality, outperforming existing baselines. Code is available at https://github.com/JamesLong199/SceneShifter.

1 Introduction

Generating audio-driven human animation with temporally controllable background scenes is important for scripted video production, virtual presentations, and social media content creation, where scene transitions must align with narration, dialogue, or editing timestamps. Recent diffusion-based video and audio-driven animation models [36, 43, 16, 9, 6] have made substantial progress in generating high-quality human videos with synchronized lip motion, expressive body movement, and realistic backgrounds.

However, existing methods struggle with **Multi-Scene Temporal Control**: generating multiple background scenes at specified timestamps while preserving a coherent foreground subject. As shown in Fig. 1, state-of-the-art image-to-video models [35] typically fail in two ways: the generated video either remains locked to the original background, or changes the scene at the cost of foreground subject consistency. Our analysis of inter-frame self-attention shows that successful multi-scene generation requires **Semantic-Aware Temporal Decoupling**, in which background and foreground

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

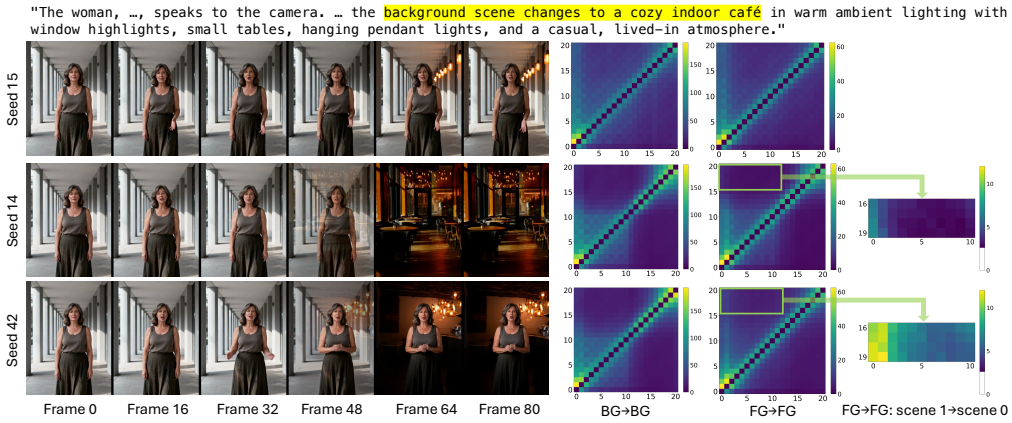
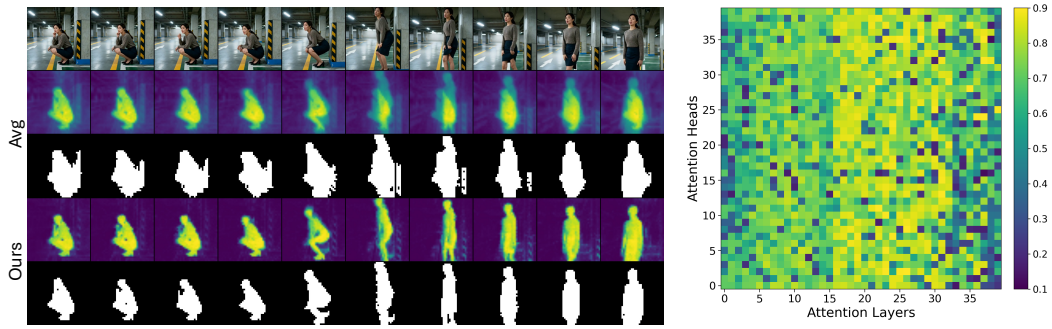


Figure 1: Videos generated from the same reference image and multi-scene prompt using different random seeds. (***Top***) An example of failing to change the background scene. (***Middle***) An example of failing to preserve the subject across scenes. (***Bottom***) An example of successful generation. The results are generated using the prompt shown above with different seeds.



(a) Reference-to-video attention heatmaps and extracted binary masks.

(b) Mask IoU heatmap.

Figure 2: Results of different foreground localization strategies on a video with large subject motion. **Avg**: attention-score averaging across all heads within each DiT layer.

tokens follow different temporal attention patterns. Background attention should be decoupled across scene segments so that each segment can form a distinct scene at the desired time. In contrast, foreground attention should remain connected across scenes so that foreground tokens in the new scene can still attend to those in the previous scene, preserving subject appearance and motion continuity. When this cross-scene foreground reference is weak, the subject may change appearance, shift position, or disappear after the transition. These observations motivate explicit self-attention guidance that isolates background attention across scenes while preserving semantic continuity among foreground tokens.

Another challenge is **Foreground Motion Localization** under large subject motion, because attention guidance relies on separating foreground from background tokens. As shown in Fig. 2a, large displacement and pose variation make accurate subject localization across frames difficult. MultiTalk [16] uses reference-to-video attention scores to localize subjects across frames by comparing each video token's average similarity to masked subject and background regions in the reference image. However, the attention scores are averaged across heads in each DiT layer, making this approach unreliable under large motion. Different layers and heads respond to different visual cues: some attend to the subject's current location, whereas others remain biased toward the reference position or background context. Averaging them therefore blurs motion-specific evidence and produces ambiguous masks. As shown by the IoU heatmap in Fig. 2b, masks from individual DiT self-attention heads differ widely in accuracy. The resulting masks can cover both the subject's current and reference positions, as illustrated in the second and third rows of Fig. 2a. In multi-scene generation, these inaccurate masks can cause semantic leakage across scenes and degrade background perceptual quality.

To address these challenges, we introduce **SceneShifter**, a training-free framework for precise multi-scene temporal control in audio-driven human animation. SceneShifter begins with spatio-temporal self-attention guidance to achieve semantic-aware temporal decoupling: background tokens are isolated across scene segments to enforce precise temporal cuts, while foreground tokens retain cross-scene attention to preserve subject appearance and motion continuity. To make this guidance robust under large subject motion, we analyze the sensitivity of self-attention to subject motion trajectories across layers and heads, and propose a fine-grained foreground motion localization method that derives accurate masks from the most informative self-attention components, as illustrated in the last two rows of Fig. 2a. To enable systematic evaluation, we establish **SceneShifterBench**, a dedicated benchmark for multi-scene audio-driven human animation that includes challenging cases with large motion and multiple subjects, along with evaluation protocols for temporal accuracy, subject preservation, and visual quality. Extensive experiments demonstrate that SceneShifter achieves accurate frame-level scene timing, robust subject consistency, and high-quality generation, substantially outperforming existing state-of-the-art methods.

In summary, we introduce **SceneShifter**, a training-free framework for multi-scene temporal control in audio-driven human animation, with the following contributions:

- We formulate multi-scene temporal control as a semantic-dependent attention control problem, where background tokens require cross-scene isolation and foreground tokens require cross-scene continuity. Based on this formulation, we propose a spatio-temporal self-attention guidance strategy for precise scene cuts and coherent subject preservation.
- We analyze the sensitivity of self-attention to subject motion trajectories across layers and heads, and propose a fine-grained **Foreground Motion Localization** method that extracts dynamic foreground masks from the most informative self-attention components for large-motion scenarios.
- We establish **SceneShifterBench**, a dedicated benchmark for evaluating multi-scene audio-driven human animation under large motion, multiple subjects, and occlusions. Experiments demonstrate strong overall performance across temporal accuracy, subject preservation, and visual quality.

2 Related Work

2.1 Audio-driven Human Animation

GAN-based [10] audio-driven human animation methods provide coarse lip synchronization with limited realism, expressiveness, and temporal stability [25, 48, 37, 38, 46, 22]. Diffusion-based methods improve visual fidelity and motion quality. Earlier approaches extend latent diffusion models [28] with 3D convolutions, audio attention, and temporal modules [32, 50, 23, 19, 21], but primarily target frontal, tightly cropped subjects and generalize poorly to complex backgrounds.

More recent approaches build on DiT-based video generators [35, 15], enabling audio-driven animation with wider shots and richer backgrounds [36, 43, 16, 9, 6]. These methods add audio conditioning to pretrained backbones but assume a single continuous background and lack precise temporal scene control. Our work enables training-free, temporally controlled background changes while remaining compatible with these pretrained models.

2.2 Temporally Controlled Video Generation

These works study how to control when text-specified events or concepts appear in a video. MinT [40] controls temporal events through temporal cross-attention and rescaled positional encoding but requires fine-tuning on temporally annotated datasets. SwitchCraft [41] steers frame-level attention toward relevant prompts without additional training, but its cross-attention control struggles with large changes such as background transitions under complex prompts. DiTCtrl [3] treats multi-prompt generation as temporal video editing, using training-free attention sharing in MM-DiT architectures [44]. It produces smooth transitions but assumes fixed-duration scenes and struggles with cross-scene subject preservation. Our training-free method supports arbitrary scene timing through self-attention decoupling and preserves subject consistency through fine-grained foreground motion localization.

2.3 Mask Guidance in Attention

Mask guidance localizes edit regions and constrains attention manipulation in diffusion-based image and video editing [30, 31]. Many methods [11, 26, 3] derive masks from target tokens' cross-attention maps, but background leakage and distractor objects can make these masks noisy [42], limiting their reliability in complex, multi-subject scenes. MultiTalk [16] instead uses reference-to-video attention to localize subjects through similarity to masked reference-image regions. However, masks computed using all attention heads can become ambiguous under large subject motion. Since our task requires finer foreground tracking for precise scene decoupling and natural subject-scene interaction, we analyze localization across self-attention layers and heads, and select informative components that accurately track subject trajectories.

3 Methods

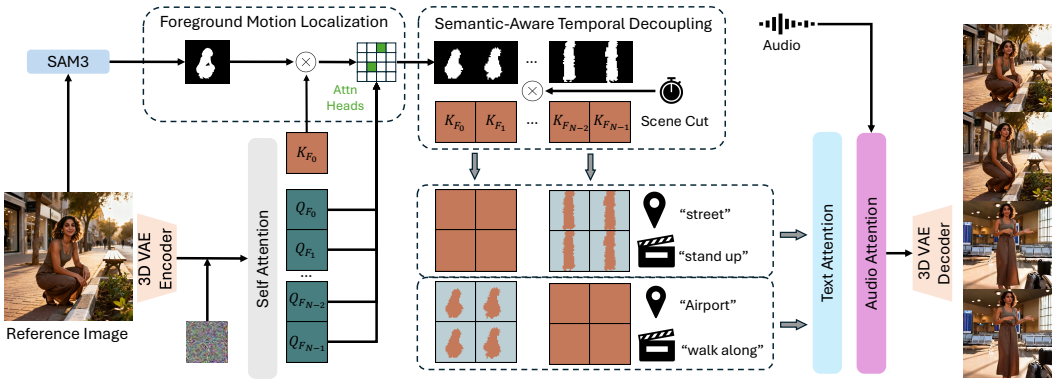


Figure 3: Pipeline of the proposed SceneShifter framework.

In this section, we present **SceneShifter**, a training-free framework for multi-scene temporal control in audio-driven human animation. As shown in Fig. 3, SceneShifter augments pretrained DiT-based video generators [16, 36] with explicit temporal control, without modifying model parameters. Given a reference image, audio, user-specified scene cut timings, and scene-specific text prompts, SceneShifter partitions video generation into scene segments. Within each DiT block, it applies self-attention guidance to enforce precise scene transitions while preserving subject consistency across scenes. To accurately separate foreground and background, SceneShifter also performs foreground motion localization for fine-grained subject tracking. Sec. 3.1 introduces the task formulation, Sec. 3.2 presents semantic-aware temporal decoupling, and Sec. 3.3 describes foreground motion localization.

3.1 Preliminaries

3.1.1 Audio-driven Human Animation

Recent audio-driven human animation methods [36, 43, 16, 9] build on powerful I2V diffusion models to generate expressive human motion and realistic backgrounds. They extend a DiT-based I2V backbone [35] with audio conditioning modules, enabling audio+image-to-video (AI2V) generation.

Given a reference image I_{ref}, speech audio A, and a text prompt T, AI2V generates a video in which the subject in I_{ref} speaks according to A and moves within a scene described by T. At inference, I_{ref} is encoded by a VAE and concatenated to the noisy video latent z_t at the start of denoising. Each DiT block then applies self-attention to model spatiotemporal dependencies, performs text cross-attention with text condition c_{text} encoded from T [27], and applies audio cross-attention with audio condition c_{audio} extracted from A [2]. Together, these attention modules integrate visual, textual, and auditory signals to produce audio-synchronized human animation.

3.1.2 Multi-Scene Temporal Control

We define **Multi-Scene Temporal Control** as changing the background scene at a user-specified time while preserving consistent foreground subject appearance and motion. To specify the transition,

we introduce a temporal condition signal f_{cut} that encodes the scene-cut timing in the latent frame space and inject f_{cut} into the model's attention mechanisms. Because the video latents are temporally downsampled, f_{cut} corresponds to a scalar index compressed from the original video frame space by a factor of 4. Although our method naturally extends to videos with more than two scenes, for clarity it is presented in the two-scene setting in the following sections, where the generated video $\mathcal{V} = \{\mathcal{V}_{S_0}, \mathcal{V}_{S_1}\}$ is divided into segments corresponding to the initial scene S_0 and the new scene S_1.

3.2 Semantic-Aware Temporal Decoupling

Semantic-aware temporal decoupling requires distinct temporal attention patterns for background (BG) and foreground (FG) tokens. BG tokens should be isolated across scenes to enforce changes at specified timestamps, while FG tokens should retain cross-scene attention to preserve subject appearance and motion continuity. We first analyze these patterns in the self-attention maps of a pretrained DiT [35], and then introduce spatio-temporal self-attention guidance to enforce them.

Self-attention analysis. We generate videos from the same reference image and multi-scene prompt using three random seeds. Fig. 1 shows three outcomes: no background change, background change with subject loss, and background change with subject preservation. We extract FG masks [4] from the generated frames and downsample them to latent resolution to separate FG and BG tokens. For seed 14, where the subject disappears, we reuse the last detected FG mask as a pseudo mask for subsequent frames to keep attention statistics comparable across seeds.

We compute inter-frame self-attention statistics over the first 50% of denoising steps, accumulating BG-to-BG and FG-to-FG scores separately for each pair of latent frames. The resulting BG→BG heatmaps in Fig. 1 associate successful background transitions with suppressed cross-scene BG attention. In the first row, strong cross-frame BG attention supports a single continuous background. Videos with scene changes instead exhibit clear attention separation at the boundary, supporting temporal BG decoupling to initiate a new background at the specified timestamp.

Foreground attention follows a different temporal pattern. To examine cross-scene foreground references, we visualize the block where FG tokens in Scene 1 query those in Scene 0, labeled FG→FG: Scene 1→Scene 0 in Fig. 1. Comparing subject-inconsistent and successful generations shows the importance of retaining attention to foreground tokens in the preceding scene. Weak or absent cross-scene FG attention can lead to subject disappearance or identity changes after the transition. These observations motivate semantic-aware temporal decoupling: suppressing cross-scene BG attention while preserving FG references.

Spatio-temporal self-attention guidance. We enforce semantic-aware temporal decoupling through spatio-temporal self-attention guidance. Each DiT block projects its input hidden features H_t into $Q = H_t W_Q$, $K = H_t W_K$, and $V = H_t W_V$, which are then partitioned at the user-specified cut frame f_{cut} into $(Q_{S_0}, K_{S_0}, V_{S_0})$ and $(Q_{S_1}, K_{S_1}, V_{S_1})$, as illustrated in Fig. 3. Per-frame foreground masks further separate each scene's keys and values into FG and BG components:

$$K_{S_0} = \{K_{S_0}^{\text{FG}}, K_{S_0}^{\text{BG}}\}, \quad V_{S_0} = \{V_{S_0}^{\text{FG}}, V_{S_0}^{\text{BG}}\}, \tag{1}$$

$$K_{S_1} = \{K_{S_1}^{\text{FG}}, K_{S_1}^{\text{BG}}\}, \quad V_{S_1} = \{V_{S_1}^{\text{FG}}, V_{S_1}^{\text{BG}}\}. \tag{2}$$

Here, braces denote concatenation along the token dimension. Each scene retains all within-scene features but excludes BG features from the other scene:

$$\omega_{S_0}^{\text{self}} = \text{Attn}(Q_{S_0}, \{K_{S_0}, K_{S_1}^{\text{FG}}\}, \{V_{S_0}, V_{S_1}^{\text{FG}}\}), \tag{3}$$

$$\omega_{S_1}^{\text{self}} = \text{Attn}(Q_{S_1}, \{K_{S_0}^{\text{FG}}, K_{S_1}\}, \{V_{S_0}^{\text{FG}}, V_{S_1}\}). \tag{4}$$

This isolates backgrounds across scenes while preserving FG references for appearance consistency and motion continuity. Each scene segment is conditioned on its corresponding prompt through text attention. Adaptation to MM-DiT architectures is detailed in Appendix B.1.

Transition context window. While cross-scene background suppression enables scene separation, it can disrupt temporal continuity near scene boundaries. To mitigate this, we introduce a short transition context window: queries in the first T_{trans} latent frames of each new scene additionally attend to both foreground and background from the last T_{ctx} latent frames of the preceding scene. This relaxation is applied only during late denoising steps to smooth transitions while preserving scene separation, with the original attention policy retained elsewhere.

3.3 Foreground Motion Localization

Semantic-aware temporal decoupling requires accurate FG-BG separation, especially under large subject motion, as shown in the first row of Fig. 2a. In image-to-video generation, reference-to-video attention can be used to localize the FG subject by measuring each video token's similarity to the FG and BG regions of the reference image [16]. The attention map between video-token queries Q_{vid} and reference-token keys K_{ref} is computed as:

$$P = \mathrm{Softmax}\left(\frac{Q_{vid}(K_{ref})^\top}{\sqrt{d}}\right), \tag{5}$$

where d is the head dimension. For each video token i, we average its attention scores over the reference FG and BG token sets Ω_{FG} and Ω_{BG} to obtain the FG and BG scores, respectively:

$$S_{FG}(i) = \frac{1}{|\Omega_{FG}|}\sum_{j\in\Omega_{FG}} P_{ij}, \quad S_{BG}(i) = \frac{1}{|\Omega_{BG}|}\sum_{j\in\Omega_{BG}} P_{ij}. \tag{6}$$

We classify token i as foreground when $S_{FG}(i) > S_{BG}(i)$, and as background otherwise.

In MultiTalk [16], the attention scores are averaged across all attention heads before computing FG and BG scores. This can produce ambiguous masks covering both the subject's reference and current positions, as shown in the second and third rows of Fig. 2a. In multi-scene generation, such masks can allow nearby background objects to leak across scenes. We attribute this ambiguity to differences in attention head sensitivity to subject motion. To examine this, we use segmentation masks [4] extracted from the generated frames as ground truth and report each layer-head pair's mean foreground and background intersection over union (IoU) in Fig. 2b. The substantial variation indicates that only a subset of heads reliably tracks subject trajectories.

We select heads with high localization accuracy and average their reference-to-video attention maps before computing FG and BG scores. For MultiTalk, we generate 20 single-scene videos using reference images from SceneShifterBench and large-motion prompts. We rank layer-head pairs by their mean foreground and background IoU against these ground-truth masks, averaged across the videos and the first 40% of denoising steps. Table 1 shows that averaging L37H32 and L37H35 outperforms the best individual head. As shown in the last two rows of Fig. 2a, the resulting masks better follow subject motion.

Table 1: Localization accuracy of individual and averaged heads. L and H denote layer and head indices; "+" denotes attention-score averaging.

Head(s)	L39H25	L37H32	L37H35	L32H3	L37H24	L37H32 + L37H35	L39H25 + L37H35
Mean IoU	0.862	0.860	0.859	0.858	0.855	**0.870**	0.869

4 Experiments

4.1 SceneShifterBench

Existing human animation datasets [48, 23] mainly feature heads or upper bodies with limited backgrounds. To address this gap, we introduce SceneShifterBench, a multimodal benchmark with full-body subjects, diverse backgrounds, and task-specific metrics for multi-scene temporal control.

Dataset. SceneShifterBench contains 150 samples, each pairing a reference image with scene-specific prompts and speech audio. The images are generated by Seedream 4.0 [29], covering diverse subject appearances, poses, backgrounds, and aspect ratios, with 40 multi-subject cases. The prompts in each sample describe up to five scenes with subject motions, while the audio is drawn from [48].

Evaluation Metrics. We evaluate temporal accuracy, human quality, and background quality. Temporal accuracy comprises **Scene Cut Success Rate (Succ.)**, the percentage of videos with the correct number of detected cuts, and **Scene Timing Accuracy (Acc.)**, the percentage of expected scene cuts correctly detected within their ground-truth frame intervals. Human quality includes

Spatial Consistency, measured by human-mask IoU across scene boundaries and averaged over videos with valid detected cuts, **Identity** and **Clothes** from VBench2 [49], and **Lip Synchronization (Sync.)** [25]. For identity, we report raw feature similarity rather than thresholded success rate. Background quality comprises VBench **Imaging Quality** and **Aesthetic Quality** [12], alongside **Scene Semantics**, evaluated only on newly generated scenes. Scene semantics sums two normalized VLM scores for background–prompt alignment and subject–environment interaction, yielding a score in $[0, 2]$, evaluated using LLaVA-Video-7B [47]. Detailed protocols are provided in Appendix A.3.

4.2 Experiment Settings

Baselines. We evaluate our method on five base models: MultiTalk [16], FantasyTalking [36], InfiniteTalk [43], WanS2V [9], and HunyuanVideo-Avatar [6]. For comparison, we integrate DiTCtrl [3] and SwitchCraft [41] into the underlying video diffusion backbones of these models, adapting their attention control mechanisms to the AI2V setting. DiTCtrl restricts attention to tokens within each scene, except at selected denoising timesteps and layers where KV sharing is enabled. During KV sharing, FG queries in Scene 1 attend only to FG keys and values in Scene 0, while BG queries in Scene 1 attend only to BG keys and values in Scene 0. In SwitchCraft, the attention modulation is activated during the first 50% of denoising steps, following the original paper. We use scene names as action words and determine event length ratios from the scene lengths in latent frames.

Implementation Details. For most backbones, we briefly delay self-attention guidance to preserve foreground motion continuity while enforcing background separation, and enable audio attention only during the latter half of denoising to limit interference with scene changes while maintaining lip synchronization, as detailed lip motion emerges during middle-to-late denoising stages [24]. For foreground motion localization, we use SAM 3 [4] to extract the human mask from the reference image and compute reference-to-video attention maps using backbone-specific sets of attention heads. All experiments are conducted on NVIDIA A800 GPUs with 80 GB of VRAM. We apply background-only query scaling adapted from [13] to mitigate background degradation in short scenes. Backbone-specific parameter settings of SceneShifter and baselines are provided in Appendix B.1.

4.3 Performance Analysis

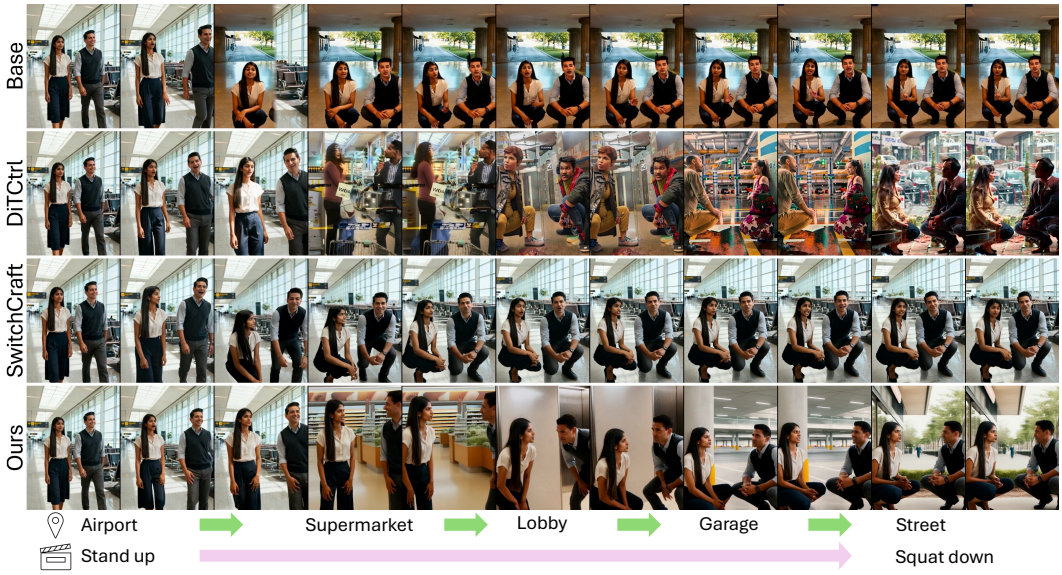


Figure 4: Qualitative results generated by the base model, DiTCtrl, SwitchCraft, and our method using the MultiTalk backbone.

Table 2 shows that SceneShifter achieves strong overall performance on SceneShifterBench, enabling accurate scene transitions while maintaining subject consistency and visual quality across five backbones. Note that WanS2V Base fails to generate any new background and all its detected cuts

Table 2: Quantitative results on the SceneShifterBench, generated using MultiTalk, FantasyTalking, InfiniteTalk, WanS2V, and HunyuanVideo-Avatar backbones. **Succ.**: scene cut success rate. **Acc.**: scene timing accuracy. **Sync.**: lip sync. Higher is better for all metrics with best scores in **bold**.

	Method	Temporal Accuracy		Human Quality				Background Quality		
		Succ.	**Acc.**	**Spatial**	**Clothes**	**Identity**	**Sync.**	**Aesthetic**	**Imaging**	**Semantics**
MultiTalk	Base	38.7%	17.5%	0.834	83.4%	0.579	**1.46**	0.541	67.5	1.61
	DiTCtrl	98.7%	96.6%	0.847	31.7%	0.407	1.23	0.523	63.3	1.55
	SwitchCraft	70.0%	30.6%	0.862	78.5%	0.577	1.36	**0.565**	67.6	1.48
	Ours	**99.3%**	**98.1%**	**0.943**	**100.0%**	**0.623**	1.45	0.542	**71.1**	**1.88**
Fantasy Talking	Base	44.7%	17.2%	0.770	90.3%	0.483	0.99	0.532	67.1	1.37
	DiTCtrl	98.7%	**91.9%**	0.845	31.2%	0.359	**1.01**	0.509	53.6	1.20
	SwitchCraft	32.0%	21.6%	0.575	54.9%	0.365	0.96	0.482	56.8	1.31
	Ours	**99.3%**	89.1%	**0.924**	**98.6%**	**0.545**	0.99	**0.541**	**70.6**	**1.74**
Infinite Talk	Base	34.0%	10.9%	0.802	77.9%	0.517	1.33	0.553	68.1	1.67
	DiTCtrl	**98.7%**	**94.4%**	0.853	33.6%	0.416	1.18	0.530	63.5	1.57
	SwitchCraft	34.0%	13.1%	0.802	78.4%	0.519	1.37	**0.556**	67.8	1.71
	Ours	98.0%	**94.4%**	**0.948**	**100.0%**	**0.628**	**1.41**	0.546	**71.7**	**1.87**
WanS2V	Base	1.3%	0.6%	0.976	**100.0%**	0.786	**1.29**	**0.629**	**75.7**	1.00
	DiTCtrl	**100.0%**	**100.0%**	0.788	12.4%	0.350	1.12	0.505	63.9	0.93
	SwitchCraft	0.7%	0.0%	0.928	**100.0%**	**0.787**	**1.29**	0.620	72.9	1.00
	Ours	**100.0%**	**100.0%**	**0.983**	**100.0%**	0.756	1.23	0.605	74.1	**1.62**
Hunyuan Video-Avatar	Base	51.3%	17.0%	0.776	66.7%	0.610	1.31	**0.540**	68.5	1.29
	DiTCtrl	**100.0%**	82.8%	0.817	32.8%	0.412	1.03	0.534	64.0	1.37
	SwitchCraft	54.7%	18.2%	0.735	62.1%	0.598	1.20	0.518	67.9	1.20
	Ours	**100.0%**	**85.3%**	**0.959**	**100.0%**	**0.639**	**1.37**	0.538	**68.9**	**1.41**

are false positives. Its high spatial consistency and visual-quality scores arise from maintaining a largely unchanged scene. SwitchCraft struggles to induce the requested scene changes, potentially reflecting the limited influence of its text cross-attention modulation in the AI2V setting. DiTCtrl enables scene changes but compromises visual quality, which may stem from excluding within-scene attention during KV sharing. These results highlight the importance of separating backgrounds across scenes while retaining sufficient attention context for coherent generation.

Fig. 4 illustrates these differences on a challenging multi-subject example involving four scene transitions and a pose change from standing to squatting. The base model fails to generate the full scene sequence and blends street and lobby elements into a single background, with visible identity drift across the transition. DiTCtrl produces substantial foreground and background artifacts and fails to preserve subject appearance, while SwitchCraft largely retains the initial airport background. In contrast, SceneShifter generates all four new scenes with distinct backgrounds while preserving subject appearance and a coherent progression from standing to squatting.

Table 3: Comparison with independent segment generation using MultiTalk on SceneShifterBench.

Method	Temporal Accuracy		Human Quality				Background Quality		
	Succ.	**Acc.**	**Spatial**	**Clothes**	**Identity**	**Sync.**	**Aesthetic**	**Imaging**	**Semantics**
Segment-wise	**99.3%**	96.9%	0.903	**100.0%**	0.532	1.26	**0.553**	70.1	1.83
SceneShifter	**99.3%**	**98.1%**	**0.943**	**100.0%**	**0.623**	**1.45**	0.542	**71.1**	**1.88**

Comparison with Independent Segment Generation. To evaluate the advantages of self-attention guidance for multi-scene temporal control, we compare SceneShifter with a sequential pipeline that generates each segment independently. At each scene boundary, a state-of-the-art open-source image model [39] replaces the background of the preceding segment's final frame, which then conditions the next segment. Table 3 reports lower identity similarity and spatial consistency for this pipeline, suggesting that repeated editing compromises subject appearance and position. In addition, spatial discontinuities and independent generation from segmented audio may disrupt audiovisual continuity and contribute to lower lip synchronization. Fig. 5 provides a qualitative comparison, showing visible



Figure 5: Comparison with independent segment generation and image-based background replacement. The prompted scene sequence is a rooftop terrace, a hospital waiting area, a subway train, a train station concourse, and a university classroom.

subject appearance changes across segments. The incomplete replacement of the train station with a classroom further illustrates the limited reliability of image-based background editing. Beyond these quality limitations, the sequential pipeline requires an additional image-editing operation at every scene boundary. The image-editing model also introduces GPU memory and model-loading requirements, complicating deployment.

4.4 Ablation Study

Foreground motion localization. We compare our selected-head localization with a fixed foreground mask from the reference image and masks estimated by averaging reference-to-video attention across all heads. Table 4 shows that our method achieves the best temporal accuracy and lip synchronization, with competitive identity, spatial consistency, and background semantics. Fixed masks cannot follow subject motion, disrupting cross-scene foreground references and weakening identity and spatial consistency. All-head averaging slightly improves subject consistency, but its broader, ambiguous masks allow background leakage and reduce semantic accuracy.

Fig. 6 compares three challenging cases with substantial subject motion. Fixed masks cause appearance changes in the first and third examples, altering the hairstyle and adding glasses, respectively. Both baseline mask strategies also cause background leakage. In the second example, the subject leaves the fixed mask region, allowing background tokens corresponding to the bench to serve as foreground references. All-head averaging produces masks that also cover the subject's reference position, causing seat leakage in the first two examples and fence discontinuities in the third. In comparison, our foreground localization better follows subject motion, preserving subject appearance while reducing background leakage.



Figure 6: Comparison of foreground localization mask strategies. **Fixed**: fixed mask from reference image. **Avg**: mask computed from reference-to-video scores averaged across all attention heads.

Transition context window. We ablate the transition context window on MultiTalk to examine its effect on visual continuity near scene boundaries. As shown in Fig. 7, background structures fluctuate across nearby frames without the window, as highlighted by the green boxes. Providing context from the preceding scene reduces these artifacts while preserving subject appearance and distinct backgrounds.

Table 4: Foreground localization ablation using MultiTalk on all 150 samples of SceneShifterBench. The transition context window is enabled for all strategies.

Method	Temporal Accuracy		Human Quality				Background Quality		
	Succ.	**Acc.**	**Spatial**	**Clothes**	**Identity**	**Sync.**	**Aesthetic**	**Imaging**	**Semantics**
Fixed	98.7%	95.0%	0.933	**100.0%**	0.608	1.41	0.543	71.1	**1.91**
Avg	98.0%	96.3%	**0.947**	**100.0%**	**0.625**	1.39	**0.545**	**71.3**	1.85
Ours	**99.3%**	**98.1%**	0.943	**100.0%**	0.623	**1.45**	0.542	71.1	1.88



Figure 7: Transition context window ablation on MultiTalk. Frames around an urban-street-to-bus-stop transition are shown in temporal order.

4.5 User Study

We compare the base model, DiTCtrl, SwitchCraft, and SceneShifter on MultiTalk using 10 randomly selected examples. Method names are concealed and result order is randomized per question. All 14 participants select the best result for each example on transition quality, subject appearance and motion consistency, and background quality, yielding 140 judgments per criterion. SceneShifter leads all three criteria in Table 5.

Table 5: User preference shares (%). Higher is better.

Criterion	Base	DiTCtrl	SwitchCraft	Ours
Transition quality	23.6	5.0	16.4	**55.0**
Subject consistency	13.6	3.6	17.1	**65.7**
Background quality	25.7	3.6	18.6	**52.1**

5 Limitations

As a training-free method, SceneShifter is subject to the capability limits of the underlying pretrained models [34] and cannot induce large movements or pose changes that they fail to generate. Despite query scaling [13], background quality can degrade when individual scenes are much shorter than the base model's training clips. In rare cases of multi-subject video generation, severe inter-subject occlusion can merge foreground masks, causing subject disappearance and appearance drift upon re-entry. Illustrative examples of these limitations are provided in Appendix E.

6 Conclusion

We present **SceneShifter**, a training-free framework for precise multi-scene temporal control in audio-driven human animation. Through **Semantic-Aware Temporal Decoupling**, it isolates background attention across scenes for accurate scene cuts and semantics while preserving cross-scene foreground attention for subject appearance and motion consistency. To handle large subject motion, **Foreground Motion Localization** selects informative self-attention components to extract dynamic foreground masks. We also introduce **SceneShifterBench**, which includes challenging large-motion and multi-subject cases. Experiments and ablations show that SceneShifter outperforms existing baselines in temporal accuracy, subject preservation, and visual quality. Future work could integrate SceneShifter into agentic video generation and editing workflows [34, 45] to support more flexible and creative applications.

7 Acknowledgements

This research / project is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

References

[1] aigc apps. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL `https://github.com/aigc-apps/VideoX-Fun`.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[3] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7763–7772, 2025.

[4] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[5] Brandon Castellano. Pyscenedetect: Video scene detection and splitting tool (version 0.6.7). `https://www.scenedetect.com/`. Open-source software.

[6] Yi Chen, Sen Liang, Zixiang Zhou, Ziyao Huang, Yifeng Ma, Junshu Tang, Qin Lin, Yuan Zhou, and Qinglin Lu. Hunyuanvideo-avatar: High-fidelity audio-driven human animation for multiple characters. *arXiv preprint arXiv:2505.20156*, 2025.

[7] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[9] Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[13] Zhiyu Jin, Xuli Shen, Bin Li, and Xiangyang Xue. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems*, 36: 70847–70860, 2023.

[14] Zonglei Jing, Zonghao Ying, Le Wang, Siyuan Liang, Xiaoqian Li, Mingchuan Zhang, Aishan Liu, Xianglong Liu, and Dacheng Tao. Cogmorph: Cognitive morphing attacks for text-to-image models. *IEEE Transactions on Dependable and Secure Computing*, 2026.

[15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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

[17] Tian Lan, Yang-Hao Zhou, Zi-Ao Ma, Fanshu Sun, Rui-Qing Sun, Junyu Luo, Rong-Cheng Tu, Heyan Huang, Chen Xu, Zhijing Wu, et al. A survey of automatic evaluation methods on text, visual and speech generations. *arXiv preprint arXiv:2506.10019*, 2025.

[18] Siyuan Liang, Jiayang Liu, Jiecheng Zhai, Tianmeng Fang, Rongcheng Tu, Aishan Liu, Xiaochun Cao, and Dacheng Tao. T2vshield: Model-agnostic jailbreak defense for text-to-video models. *International Journal of Computer Vision*, 134(4):144, 2026.

[19] Gaojie Lin, Jianwen Jiang, Chao Liang, Tianyun Zhong, Jiaqi Yang, and Yanbo Zheng. Cyberhost: Taming audio-driven avatar diffusion model with region codebook attention. *arXiv preprint arXiv:2409.01876*, 2024.

[20] Jiayang Liu, Siyuan Liang, Shiqian Zhao, Rongcheng Tu, Wenbo Zhou, Xiaochun Cao, Dacheng Tao, and Siew Kei Lam. Jailbreaking the text-to-video generative models. *arXiv e-prints*, pages arXiv–2505, 2025.

[21] Jianzhi Long, Wenhao Sun, Rong-Cheng Tu, and Dacheng Tao. Lightning fast caching-based parallel denoising prediction for accelerating talking head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7503–7511, 2026.

[22] Jianzhi Long, Rong-Cheng Tu, and Dacheng Tao. Simdem: Audio-driven emotional talking head generation with simplified and decoupled expression modeling. *IEEE Transactions on Multimedia*, 2026.

[23] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5489–5498, 2025.

[24] Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.

[25] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[26] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[29] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

[30] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

[31] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. Diffusion model-based video editing: A survey, 2024. *URL https://arxiv. org/abs/2407.07111*.

[32] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

[33] Rong-Cheng Tu, Zi-Ao Ma, Tian Lan, Yuehao Zhao, He-Yan Huang, and Xian-Ling Mao. Automatic evaluation for text-to-image generation: Task-decomposed framework, distilled training, and meta-evaluation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22340–22361, 2025.

[34] Rong-Cheng Tu, Wenhao Sun, Zhao Jin, Jingyi Liao, Jiaxing Huang, and Dacheng Tao. Spagent: Adaptive task decomposition and model selection for general video generation and editing. *IEEE Transactions on Image Processing*, 2026.

[35] 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.

[36] Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9891–9900, 2025.

[37] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv*, 2021.

[38] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022.

[39] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[40] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23989–24000, 2025.

[41] Qianxun Xu, Chenxi Song, Yujun Cai, and Chi Zhang. Switchcraft: Training-free multi-event video generation with attention controls. *arXiv preprint arXiv:2602.23956*, 2026.

[42] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023.

[43] Shaoshu Yang, Zhe Kong, Feng Gao, Meng Cheng, Xiangyu Liu, Yong Zhang, Zhuoliang Kang, Wenhan Luo, Xunliang Cai, Ran He, et al. Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. *arXiv preprint arXiv:2508.14033*, 2025.

[44] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pages 83048–83077, 2025.

[45] Huanjin Yao, Ruifei Zhang, Jiaxing Huang, Jingyi Zhang, Yibo Wang, Bo Fang, Ruolin Zhu, Yongcheng Jing, Shunyu Liu, Guanbin Li, et al. A survey on agentic multimodal large language models. *arXiv preprint arXiv:2510.10991*, 2025.

[46] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023.

[47] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

[48] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021.

[49] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[50] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions: the SceneShifter framework with semantic-aware temporal decoupling and foreground motion localization, as well as SceneShifterBench. The experimental section evaluates the claimed improvements in temporal accuracy, human quality, and background quality.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Sec. 5

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results, theorems, or formal proofs. The method is an empirical training-free attention guidance framework.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the task setting, benchmark construction, baselines, evaluation metrics, attention guidance formulation, foreground localization procedure, and implementation details needed to reproduce the main experiments. The method is training-free and operates on pretrained AI2V backbones.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide our code and demo examples via the link at the end of the abstract.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the benchmark size, prompt construction, scene timing sampling, evaluation metrics, baselines, pretrained backbones, attention guidance timesteps, foreground localization setting, and GPU type. Since SceneShifter is training-free, no training is required.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We use the same seed for generation across all generation settings to ensure fair comparison because the motion trajectory and new scene background formation vary significantly across seeds.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mentioned that all experiments are done on NVIDIA A800 GPUs with 80GB VRAM.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing pretrained models and synthesized images in benchmark dataset for research purposes, and the submission is anonymized. We do not collect private user data or conduct human-subject experiments.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction discusses positive applications such as television production, virtual presentations, and social media content creation. The impact statement in the appendix discusses potential risks of generative models, including biased or harmful content, privacy violations, and intellectual property concerns.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained generative model or scraped dataset. SceneShifter is a training-free inference framework built on existing pretrained video generators; any release of code or benchmark assets will follow the usage terms of the underlying assets.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the existing models, datasets, and evaluation tools used, including pretrained AI2V backbones, SAM 3, HDTF, VBench/VBench2, and LLaVA-Video, which are open-sourced software for research usage.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces SceneShifterBench and documents its construction, including reference image generation, prompt design, scene timing sampling, challenging cases, and evaluation metrics. Additional documentation can be provided with the released benchmark assets.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Evaluation is performed using automated metrics and VLM-based questionnaires.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject studies, user studies, or crowdsourcing. Therefore, IRB approval or equivalent review is not applicable.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper uses LLaVA-Video-7B as part of the evaluation protocol for scene semantics and subject-scene interaction naturalness. This usage is described in the evaluation metrics section and does not constitute a core component of the proposed generation method.

A SceneShifterBench Details

A.1 Reference Images and Scene Prompts

SceneShifterBench contains 150 samples, each comprising a reference image, scene-specific text prompts, speech audio, and prescribed scene-transition timings. We construct image-generation prompts by sampling subject count, appearance, pose, background, and aspect ratio, and synthesize the reference images using Seedream 4.0 [29]. Each image is paired with speech audio from HDTF [48]. The reference images include 110 single-subject and 40 two-subject examples, covering 190 subjects in total. Subject attributes span nine ethnicity categories, five age groups, and five body types. Both the reference images and scene prompts cover 25 background types, including indoor and outdoor environments.

Table 6 summarizes the benchmark composition, while Tables 7 and 8 provide detailed distributions. Subject attributes describe the categories specified during synthetic image construction. Image-level statistics use 150 reference images as the denominator; subject-level statistics use 190 subjects.

Table 6: Composition of SceneShifterBench. Scene counts include the initial scene.

Attribute	Count
Reference images	
Total images	150
Single-subject / two-subject	110 / 40
Total subjects	190
Ethnicity / age / body-type categories	9 / 5 / 5
Background types	25
Aspect ratios	5
Scene prompts	
Two / three / four / five scenes	50 / 50 / 30 / 20
Total scene instances	470
Background types	25

Table 7: Subject attributes and aspect ratios. Counts cover 190 subjects or 150 images for aspect ratios.

Attribute	Category (count)
Gender	Man (95), woman (95)
Ethnicity	South Asian (24), Hispanic/Latino (23), Indigenous Australian (23), East Asian (22), mixed-ethnicity (21), Caucasian (21), Middle Eastern (21), Southeast Asian (19), Black (16)
Age	20s (47), 30s (52), 40s (31), 50s (42), 60s+ (18)
Body type	Average (42), slim (40), curvy (38), stocky (37), athletic (33)
Reference pose	Sitting (88), standing (76), squatting (26)
Aspect ratio	4:3 (35), 2:3 (30), 1:1 (29), 3:2 (28), 3:4 (28)

Table 8: Background distribution in reference images and prompted scenes, including initial scenes.

Background	Reference images	Prompted scenes
Covered walkway	8	19
Museum	8	17
Park	7	20
Bookstore	7	17
Rooftop terrace	7	21
Supermarket	7	18
Library	7	17
Urban street	7	20
Elevator lobby	7	17
Train station	6	18
Hotel lobby	6	18
Conference hall	6	20
Parking garage	6	20
Courtyard	6	17
Bus stop	6	22
Beach promenade	6	20
Café	6	21
Hospital waiting area	5	19
Airport	5	19
Classroom	5	18
Workshop	5	21
Office	5	18
Recording studio	4	19
Podcast room	4	17
Subway	4	17
Total	150	470

A.2 Evaluation Metrics

This section details the evaluation metrics introduced in Sec. 4.1. We report Succ. and Acc. as percentages, and Spatial Consistency and Identity as raw scores. Higher values indicate better performance for all four metrics.

Scene Cut Success Rate. Let $\mathcal{V}$ denote the evaluated videos with valid scene-transition metadata. For video v, let K_v be the expected number of cuts and $\hat{\mathbf{c}}_v = (\hat{c}_{v,1}, \ldots, \hat{c}_{v,\hat{K}_v})$ the sorted scene-cut indices returned by our PySceneDetect-based pipeline [5]. A video is successful when the detected

and expected cut counts match:

$$u_v = \mathbf{1}[\hat{K}_v = K_v], \qquad \text{Succ.} = \frac{100}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} u_v. \tag{7}$$

A failed detection is treated as producing zero cuts. This metric measures cut-count agreement, independently of timing, and remains subject to false-positive and missed detections.

Scene Timing Accuracy. Scene Timing Accuracy measures the percentage of all expected cuts that are correctly timed in videos with matching detected and expected cut counts. Detected cuts and expected transitions are paired in chronological order. Let $I_{v,j}$ denote the accepted video-frame interval for expected transition j.

For frame-split evaluation, let $s_{v,j}$ be its latent-frame split index. With temporal compression ratio $R = 4$, the inclusive interval is

$$I_{v,j} = [Rs_{v,j} - (R-1),\ Rs_{v,j}]. \tag{8}$$

For single-prompt evaluation, expected transitions are extracted in prompt order and assigned to early, middle, or late temporal buckets. In our experiments, Wan-based models [16, 36, 43, 9] generate 81-frame videos, while HunyuanVideo-Avatar [6] generates 129-frame videos. The corresponding inclusive intervals are:

Video length	early	middle	late
81	$[0, 26]$	$[27, 53]$	$[54, 80]$
129	$[0, 42]$	$[43, 85]$	$[86, 128]$

. (9)

Thus, the two conditioning settings use different timing tolerances but share the same per-cut scoring rule.

The number of correctly timed cuts in video v is

$$a_v = \begin{cases} \sum_{j=1}^{K_v} \mathbf{1}[\hat{c}_{v,j} \in I_{v,j}], & \hat{K}_v = K_v, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The dataset-level score is

$$\text{Acc.} = 100\, \frac{\sum_{v \in \mathcal{V}} a_v}{\sum_{v \in \mathcal{V}} K_v}. \tag{11}$$

Videos with missing or extra detected cuts receive zero timing credit, while their expected cuts remain in the denominator. For videos with matching counts, each correctly timed cut contributes independently.

Spatial Consistency. Spatial Consistency measures subject-position continuity around detected scene cuts. We use the cuts returned by the hybrid detector, regardless of whether their count or timing matches the ground truth. Videos with unsuccessful detector outputs and cuts at or beyond the last video frame are excluded. Let $M_v(t) \in \{0,1\}^{H \times W}$ denote the human mask extracted using SAM 3 [4] at frame t. For multi-subject videos, we extract each subject's mask separately and take their union to obtain $M_v(t)$.

For two binary masks M and N, we define

$$\text{IoU}(M, N) = \begin{cases} \dfrac{|M \cap N|}{|M \cup N|}, & |M \cup N| > 0, \\ 1, & |M \cup N| = 0. \end{cases} \tag{12}$$

If no person is detected in either frame of a compared pair, its IoU score is set to zero. For each detected cut $\hat{c}_{v,k}$, we use a window radius of $w = 1$ and compare adjacent-frame masks. Let $\mathcal{T}_{v,k}$ contain the frame indices within this window for which both t and $t-1$ are valid:

$$\mathcal{T}_{v,k} = \left\{ t \in \mathbb{Z} : \hat{c}_{v,k} - 1 \le t \le \hat{c}_{v,k} + 1,\ t, t-1 \text{ are valid frame indices} \right\}. \tag{13}$$

The per-cut score is

$$S_{v,k} = \frac{1}{|\mathcal{T}_{v,k}|} \sum_{t \in \mathcal{T}_{v,k}} \mathrm{IoU}\big(M_v(t), M_v(t-1)\big). \tag{14}$$

Cuts with no valid frame pairs are excluded. Let $\mathcal{K}_v$ denote the remaining cuts in video v, and $\mathcal{V}_{\mathrm{sp}}$ the videos with at least one valid cut score. We average over cuts within each video and then equally over eligible videos:

$$\mathrm{Spatial} = \frac{1}{|\mathcal{V}_{\mathrm{sp}}|} \sum_{v \in \mathcal{V}_{\mathrm{sp}}} \frac{1}{|\mathcal{K}_v|} \sum_{k \in \mathcal{K}_v} S_{v,k}. \tag{15}$$

Identity. We evaluate identity preservation using ArcFace [8]. Unlike the thresholded success rate in VBench2 [49], we report raw cosine similarity. The input reference image is resized to match the first generated frame when their dimensions differ. For each evaluated frame, we retain up to five detected faces and select the assignment maximizing similarity to the reference subjects.

Let $f_{v,r}^{\mathrm{ref}}$ denote the reference embedding for subject r and $f_{v,r}(t)$ its assigned embedding in generated frame t. For the set $\mathcal{F}_{v,r}$ of frames with valid assignments, the subject-level score is

$$\mathrm{ID}_{v,r} = \frac{1}{|\mathcal{F}_{v,r}|} \sum_{t \in \mathcal{F}_{v,r}} \frac{f_{v,r}(t)^\top f_{v,r}^{\mathrm{ref}}}{\|f_{v,r}(t)\|_2 \|f_{v,r}^{\mathrm{ref}}\|_2 + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{16}$$

A subject requires at least 20 valid frames to be included. For two-subject videos, the score averages the two subjects' identity similarities. The reported Identity score is averaged over videos with valid scores.

A.3 Scene Semantics

Scene Semantics evaluates background–prompt alignment and subject–environment interaction using LLaVA-Video-7B [47], leveraging the semantic understanding and reasoning capabilities of Vision-Language Models (VLM) [33, 17]. We construct a video question-answering questionnaire from each scene's ground-truth prompt. Evaluation includes only videos whose detected cut count matches the expected count, with scene boundaries defined by the detected cuts. We exclude the opening scene and evaluate only the newly generated scenes. The reported score sums the two component scores, each in $[0, 1]$, yielding a total in $[0, 2]$.

Background–prompt alignment. For each evaluated scene, we parse its prompt into at most five statements describing the background location and objects. The location statement takes the form "The background location is `<location>`." Additional statements take the form "The background contains `<object>`." and are extracted from background objects listed after the location clause. Object phrases are retained only if they contain a word from a fixed place-and-object vocabulary, such as bench, chair, table, window, floor, or street. If no usable statement is extracted, we use "The background scene matches the described setting." The VLM judges whether each statement is true, considering only the environment.

Subject–environment interaction. We first ask the VLM to classify the subject's pose as standing, sitting, or squatting, then evaluate pose-specific physical plausibility. Prompts describing walking are assigned the standing category. Standing and squatting each have one follow-up plausibility question. Sitting has two sequential checks: whether the subject sits on the expected support and whether the body is naturally integrated with that support. If the predicted pose does not match the prompted pose, all follow-up checks receive zero. For sitting, a failed support check also assigns zero to the integration check. Questions assigned zero by these rules are not sent to the VLM. For multi-action videos, interaction is evaluated only on the last evaluated scene.

Pose questions require a single-word answer, while the remaining questions require yes/no answers. Where requested, a negative answer includes an explanation of at most one sentence. These explanations do not contribute to the score.

Questionnaire templates. The following templates specify the evaluated questions.

Background–prompt alignment. For each parsed statement `<gt_item>`, we ask:

Please answer yes or no only for the following question. IMPORTANT: Focus only on the environment/background. Ignore the person/subject, faces, clothing, and actions. Is this statement true in the background? '<gt_item>'. Answer with yes or no only.

Subject–environment interaction. Here, <subject> is the subject phrase extracted from the prompt, such as "the woman" or "the man and the woman."

Pose classification:

What pose is <subject> performing? Choose exactly one word: standing, sitting, or squatting.

Standing plausibility:

Does the standing by <subject> look physically plausible (feet grounded, no obvious floating, no severe body–environment intersection)? Answer yes/no. If no, briefly explain ($\leq$ 1 sentence).

Squatting plausibility:

Does the squatting by <subject> look physically plausible and stably grounded (feet contact the ground plausibly; no obvious hovering, sliding, or severe ground intersection)? Answer yes/no. If no, briefly explain ($\leq$ 1 sentence).

Sitting support:

Is <subject> sitting on a '<expected_support>' (as intended by the prompt)? Answer yes or no. If no, briefly explain ($\leq$ 1 sentence).

Sitting integration:

Does the sitting by <subject> look naturally supported by the '<expected_support>' (hips/thighs contact the seat surface; body not floating above or sinking into it)? Answer yes/no. If no, briefly explain ($\leq$ 1 sentence).

If the prompt specifies no support object, the sitting questions refer to a plausible seat or surface.

Example: standing. For video 001, the evaluated scene is an indoor recording studio.

Scene prompt:

The man and the woman walk along the studio floor in indoor recording studio, soundproof walls, studio monitors, and a microphone stand near the subject, speaking to the camera.

Background–prompt alignment:

Is this statement true in the background? 'The background location is indoor recording studio.' Answer with yes or no only.

Subject–environment interaction:

What pose are the man and the woman performing? Choose exactly one word: standing, sitting, or squatting.

Does the standing by the man and the woman look physically plausible (feet grounded, no obvious floating, no severe body–environment intersection)? Answer yes/no. If no, briefly explain ($\leq$ 1 sentence).

Example: sitting. For video 003, the evaluated scene is an indoor podcast room.

Scene prompt:

The woman continues to speak to the camera while sitting on studio chair at the recording table in indoor podcast room, acoustic panels, a small recording table, and neatly arranged audio gear.

Background–prompt alignment:

Is this statement true in the background? 'The background location is indoor podcast room.' Answer with yes or no only.

Subject–environment interaction:

What pose is the woman performing? Choose exactly one word: standing, sitting, or squatting.

Is the woman sitting on a 'chair'? Answer yes or no. If no, briefly explain (≤ 1 sentence).

Does the sitting by the woman look naturally supported by the 'chair'? Answer yes/no. If no, briefly explain (≤ 1 sentence).

Component scores. For an evaluated scene σ with N_σ background statements, let $y_{\sigma,i} \in \{0,1\}$ indicate whether the VLM judges statement i to be true. The background–prompt alignment score is

$$S_{\text{align}}(\sigma) = \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} y_{\sigma,i}. \tag{17}$$

For interaction, let p_σ and $p_\sigma^\star$ denote the predicted and expected poses, respectively, and define $c_{\text{pose}}(\sigma) = \mathbf{1}[p_\sigma = p_\sigma^\star]$. For standing or squatting, let $c_{\text{plaus}}(\sigma)$ be the binary outcome of the plausibility check. For sitting, let $c_{\text{sup}}(\sigma)$ and $c_{\text{integ}}(\sigma)$ be the support and integration outcomes. The interaction score is

$$S_{\text{int}}(\sigma) = \begin{cases} \dfrac{c_{\text{pose}}(\sigma) + c_{\text{plaus}}(\sigma)}{2}, & p_\sigma^\star \in \{\text{standing}, \text{squatting}\}, \\ \dfrac{c_{\text{pose}}(\sigma) + c_{\text{sup}}(\sigma) + c_{\text{integ}}(\sigma)}{3}, & p_\sigma^\star = \text{sitting}. \end{cases} \tag{18}$$

All follow-up outcomes are zero when $c_{\text{pose}}(\sigma) = 0$. For sitting, $c_{\text{integ}}(\sigma) = 0$ whenever $c_{\text{sup}}(\sigma) = 0$. Skipped checks remain in the denominator.

Aggregation. Let $\mathcal{G}_v$ denote the evaluated scenes in video v, and $\mathcal{G}_v^{\text{int}} \subseteq \mathcal{G}_v$ the scenes evaluated for interaction. For multi-action videos, $\mathcal{G}_v^{\text{int}}$ contains only the last evaluated scene. We compute the per-video component scores as

$$\mu_{\text{align}}(v) = \frac{1}{|\mathcal{G}_v|} \sum_{\sigma \in \mathcal{G}_v} S_{\text{align}}(\sigma), \tag{19}$$

$$\mu_{\text{int}}(v) = \frac{1}{|\mathcal{G}_v^{\text{int}}|} \sum_{\sigma \in \mathcal{G}_v^{\text{int}}} S_{\text{int}}(\sigma). \tag{20}$$

Let $\mathcal{V}_{\text{sem}}$ denote the eligible evaluated videos. Each component is averaged equally across videos:

$$\mu_{\text{align}} = \frac{1}{|\mathcal{V}_{\text{sem}}|} \sum_{v \in \mathcal{V}_{\text{sem}}} \mu_{\text{align}}(v), \tag{21}$$

$$\mu_{\text{int}} = \frac{1}{|\mathcal{V}_{\text{sem}}|} \sum_{v \in \mathcal{V}_{\text{sem}}} \mu_{\text{int}}(v). \tag{22}$$

The reported Scene Semantics score is

$$\text{Semantics} = \mu_{\text{align}} + \mu_{\text{int}} \in [0, 2]. \tag{23}$$

B Experimental Settings

B.1 Backbone-Specific Settings

We implement SceneShifter on MultiTalk [16], FantasyTalking [36], InfiniteTalk [43], WanS2V [9], and HunyuanVideo-Avatar [6]. Table 9 summarizes the sampling settings, attention schedules, foreground localization heads, and transition context windows for each backbone. We average reference-to-video attention scores over the listed heads for mask extraction and update the masks during the specified denoising steps. Self-attention guidance, audio conditioning, and transition smoothing follow backbone-specific schedules. All step indices are zero-based and follow denoising execution order; ranges include both endpoints. Experiments are conducted on NVIDIA A800 GPUs with 80 GB of VRAM.

Table 9: Backbone-specific settings. Denoising steps are indexed from zero in execution order, and ranges are inclusive. Transition and context lengths are measured in latent frames.

Setting	MultiTalk	FantasyTalking	InfiniteTalk	WanS2V	HunyuanVideo-Avatar
Denoising steps	40	30	40	40	50
Output frames	81	81	81	81	129
Self-attention guidance start step	3	2	3	3	0
Audio attention start step	20	15	20	0	0
Localization heads (layer, head)	(37,32), (37,35)	(37,32), (25,26), (39,25)	(37,32), (37,35)	(37,32), (37,35)	(19,22)
Mask update steps	0–14	0–8	0–14	0–2	0–14
Transition length T_{trans}	2	2	2	2	2
Context length T_{ctx}	1	1	1	1	2
Transition window activation steps	30–39	22–29	30–39	30–39	20–49

Adaptation to HunyuanVideo-Avatar. For the MM-DiT architecture in HunyuanVideo-Avatar, we apply semantic-aware temporal decoupling and foreground localization to vision-to-vision attention. Each scene prompt is independently encoded using the model's text encoder, and the resulting tokens are concatenated into the text stream. To prevent cross-scene semantic leakage, video and text queries associated with one scene cannot attend to text keys from other scenes.

Baseline Settings. For each backbone, the base model, SwitchCraft [41], and DiTCtrl [3] use the same audio attention start step as SceneShifter (Table 9). DiTCtrl activates KV sharing only in the final 20% of DiT layers during an early denoising window: steps 5–14 for MultiTalk, InfiniteTalk, and WanS2V; 2–14 for FantasyTalking; and 0–14 for HunyuanVideo-Avatar. All other layers and steps retain within-scene attention.

C Additional Quantitative Results

C.1 Robustness Across Random Seeds

We use seed 43 for all samples in the main experiments. To assess sensitivity to random initialization, we repeat the evaluation on the full SceneShifterBench using MultiTalk [16] with seeds 28 and 56, keeping all other settings fixed. Table 10 reports individual scores and their mean and sample standard deviation across seeds. Performance remains stable across temporal accuracy, human quality, and background quality, with seed 43 showing no consistent advantage. These results suggest that SceneShifter's performance is not dependent on a particularly favorable random seed.

Table 10: Random seed sensitivity on SceneShifterBench. We report means and sample standard deviations across three seed-level aggregate scores.

Seed	Temporal Accuracy		Human Quality				Background Quality		
	Succ. (%)	**Acc. (%)**	**Spatial**	**Clothes (%)**	**Identity**	**Sync.**	**Aesthetic**	**Imaging**	**Semantics**
43	99.3	98.1	0.943	100.0	0.623	1.45	0.542	71.1	1.88
28	100.0	97.8	0.947	100.0	0.620	1.48	0.546	68.4	1.82
56	99.3	96.6	0.946	100.0	0.612	1.48	0.542	70.3	1.87
Mean $\pm$ Std	99.6 ± 0.4	97.5 ± 0.8	0.945 ± 0.002	100.0 ± 0.0	0.618 ± 0.006	1.47 ± 0.02	0.543 ± 0.002	69.9 ± 1.4	1.86 ± 0.03



Figure 8: Qualitative comparison on FantasyTalking. The initial prompt specifies a squatting action; subsequent prompts specify a rooftop terrace, a covered outdoor walkway, and an indoor supermarket.

D Additional Qualitative Results

D.1 Results Across AI2V Backbones

Figs. 8–11 compare SceneShifter with the base model, DiTCtrl [3], and SwitchCraft [41] on four additional AI2V backbones, including FantasyTalking [36], InfiniteTalk [43], WanS2V [9], and HunyuanVideo-Avatar [6]. The base models and SwitchCraft often retain the initial background or fail to distinguish successive scenes, while DiTCtrl introduces appearance changes or substantial visual artifacts. SceneShifter produces more distinct scene sequences while better preserving the foreground subjects.

D.2 Extensions to Other Generation Settings

Video-driven animation. Fig. 12 demonstrates SceneShifter on Wan-Animate [7], where subject motion is controlled by a driving video. The subject remains visually consistent across changing backgrounds while performing different gestures, illustrating compatibility with video-driven human animation.

Camera-controlled generation. Fig. 13 demonstrates SceneShifter on Wan2.1-Fun-Camera-Control [1] with pan-left, zoom-out, and clockwise camera trajectories. These examples illustrate how SceneShifter can be combined with complementary controls to enable more creative video generation.

Longer videos. Fig. 14 presents a longer 3-clip video generated using MultiTalk [16] as backbone, with five motion frames used between consecutive clips. The example maintains two recognizable subjects through multiple background changes, extending the qualitative evaluation beyond a single short clip.



Figure 9: Qualitative comparison on InfiniteTalk. The prompted scene sequence is a covered outdoor walkway, a rooftop terrace, and a beachside promenade.



Figure 10: Qualitative comparison on WanS2V. The prompted scene sequence is a hospital waiting area, an office, a subway train, and a hotel lobby lounge.

E Limitations

E.1 Limitations of Base-Model Motion Generation

SceneShifter relies on the pretrained backbone to synthesize the requested subject motion. In Fig. 15, the WanS2V-based result [9] changes the background but retains the subject's seated pose, whereas the MultiTalk-based result [16] produces the requested standing motion. Thus, successful scene control does not guarantee execution of motion instructions that the underlying model struggles to follow.

E.2 Background Degradation in Short Scenes

Background quality can degrade when scene segments are much shorter than the base model's training clip length. Fig. 16 shows severe background artifacts without query scaling. Applying query scaling, following [13], substantially reduces these artifacts in the illustrated example. This mitigation improves short-scene generation but does not eliminate the dependence on the base model's capabilities.



Figure 11: Qualitative comparison on HunyuanVideo-Avatar. The prompted scene sequence is a bookstore, a university classroom, and a recording studio.



Figure 12: Video-driven animation with SceneShifter on Wan-Animate. The prompted scene sequence is a supermarket, a conference hall, and a parking garage. Background changes are accompanied by subject gestures controlled by the driving video.



Figure 13: SceneShifter with camera control on Wan2.1-Fun-Camera-Control. Top: pan left, with a prompted transition from a park to a hospital waiting area. Middle: zoom out, from a museum gallery to a parking garage. Bottom: clockwise camera motion, with a prompted sequence of a supermarket, a conference hall, an urban street, and a museum gallery.



Figure 14: Longer video generation with SceneShifter on MultiTalk. The prompted scene sequence is a parking garage, a library study area, a bus stop, a conference hall, an urban street. Selected frames span three generated clips and show preservation of both subjects across multiple scene changes.

E.3 Severe Inter-Subject Occlusion

Under severe inter-subject occlusion, foreground masks can merge, causing a subject to disappear after a scene transition. The resulting interruption of cross-scene foreground references can lead to



Figure 15: Dependence on the base model's motion-generation capability. Both rows apply SceneShifter to the prompted sequence of an urban street, a bus stop, a hotel lobby lounge, a park, and a café. WanS2V (top) retains the seated pose despite the standing instruction, while MultiTalk (bottom) generates the pose change.



Figure 16: Effect of query scaling on short-scene generation with MultiTalk. The prompted scene sequence is an urban street, a bus stop, a hotel lobby lounge, a park, and a café. Without query scaling (top), later scenes exhibit severe background artifacts; query scaling (bottom) substantially improves their visual quality.

appearance drift upon re-entry. Fig. 17 illustrates subject disappearance and subsequent changes in appearance. Preserving subject-specific references through heavy occlusion and temporary absence remains a direction for future work.



Figure 17: Failure case involving subject disappearance and re-entry. One of the two initial subjects disappears during the sequence, and their appearances differ substantially in later frames. The example illustrates the difficulty of maintaining subject identity when foreground continuity is interrupted. Frames are shown in temporal order.

E.4 Limitations of Automated Evaluation

SceneShifterBench relies on automated scene-cut detection using PySceneDetect [5], which can mistake camera motion or visual artifacts for scene transitions and miss gradual or subtle transitions. These false positives and false negatives affect Scene Cut Success Rate and Scene Timing Accuracy. Detection errors also influence Spatial Consistency, for which subject masks are compared at detected cuts, and Scene Semantics, for which detected cuts define scene segments and only videos satisfying the Scene Cut Success Rate criterion are included. In future work, we will explore VLM-assisted scene detection to better distinguish semantic scene changes.

F Societal Impact

SceneShifter can support creative video production, but may also be misused to create misleading depictions of individuals or harmful content. Users could manipulate reference images to embed harmful contextual elements [14] or inject malicious prompts into video generation through jailbreak attacks [20]. Responsible deployment should prioritize consent, clear disclosure of synthetic content, and safeguards against malicious inputs [18].