SocialDirector: Training-Free Social Interaction Control for Multi-Person Video Generation

Liangyang Ouyang[1] **Ruicong Liu**[1,2*] **Caixin Kang**[1] **Yifei Huang**[1,2] **Yoichi Sato**[1]
[1]The University of Tokyo [2]AlayaLab
{oyly, lruicong, cxkang, hyf, ysato}@iis.u-tokyo.ac.jp



Figure 1: We propose **SocialDirector**, a training-free controller that enhances multi-person video generation with explicit control over social interactions. Based on a pretrained image-to-video diffusion transformer, SocialDirector controls *who* performs *what* action, *when* each action occurs, and *toward whom* it is directed, producing faithful interactions while preserving video quality.

Abstract

Video generation has advanced rapidly, producing photorealistic videos from text or image prompts. Meanwhile, film production and social robotics increasingly demand *multi-person* videos with rich social interactions, including conversations, gestures, and coordinated actions. However, existing models offer no explicit control over interactions, such as *who* performs *which* action, *when* it occurs, and *toward whom* it is directed. This often results in wrong person performing unintended actions (actor-action mismatch), disordered social dynamics, and wrong action targets. To address these challenges, we present **SocialDirector**, a training-free interaction controller that enhances the generation model by modulating cross-attention maps. SocialDirector contains two modules: *Social Actor Masking* and *Directional Reweighting*. Social Actor Masking constrains each person's visual tokens to attend only to their own textual descriptions via a spatiotemporal mask, avoiding actor-action mismatch and disordered social dynamics.

*Corresponding author.

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

Directional Reweighting amplifies attention to directional words (e.g., "leftward", "right"), leading each action towards its intended target. To evaluate generated social interactions, we annotate existing datasets with interaction descriptions and build a fully automated evaluation pipeline powered by open-source VLMs. Experiments on different video generation models show that SocialDirector significantly improves interaction fidelity and approaches the upper bound set by real videos. Code is available at `https://github.com/ut-vision/SocialDirector`.

1 Introduction

Diffusion-based video generation has advanced rapidly in recent years. Large-scale models such as Seedance [16], Wan [62], and LTX [19] can now synthesize photorealistic characters, motions, and camera movements from text or image prompts. Building on this progress, studies have started extending to multi-person scenarios [29, 41, 21, 88, 64], motivated by applications in film production [55, 76] and social robotics [7, 4]. Among these, social interactions, where individuals engage in conversations, gestures, and coordinated actions, have drawn growing attention. Faithfully generating such interactions is essential for delivering immersive experiences.

However, existing generative models lack explicit control over interactions, such as *who* performs *what* action, *when* it occurs, and *toward whom* it is directed. Consequently, generated videos often exhibit wrong person performing unintended actions (actor-action mismatch), disordered social dynamics, and wrong action targets. We attribute this failure to *attention leakage* in the cross-modal attention process: in a standard diffusion transformer (DiT) block, every visual token attends to all text tokens through cross-attention. As a result, the model imposes no spatial, temporal, or relational priors that tie a described action to its intended actor, moment, and target.

To address this limitation, we propose **SocialDirector**, an interaction controller that enhance the generation model by modulating its cross-attention maps. SocialDirector contains two training-free modules: *Social Actor Masking* and *Directional Reweighting*. Social Actor Masking takes as input the first-frame image, per-person bounding boxes, and text prompt with timestamps. From these, it constructs a spatiotemporal mask that restricts each person's visual tokens to attend only to their own textual description within the relevant time window. This design eliminates attention leakage across person and time, avoiding actor-action mismatch and disordered social dynamics. Directional Reweighting further amplifies attention to directional words (e.g., "left", "right"), leading each action towards its intended target. This module sharpens the model's sensitivity to directional cues, reducing mistargeted actions. SocialDirector introduces no trainable parameters or architectural modifications, and can be seamlessly plugged into DiT-based video generation models.

Evaluating generated social interaction is an open challenge. Existing video quality metrics only measure visual quality, not whether actor, action, and target are correctly bound. To this end, we annotate 149 samples from three existing multi-person datasets (MELD [54], MMSI [32], and SocialGesture [5]) with structured interaction descriptions. Each description explicitly specifies the actor, action, target, timestamp, and per-person bounding boxes. Based on these annotations, we build the first fully automated evaluation pipeline for social interactions. This pipeline formulates the evaluation as binary video question answering, with answers aggregated by majority vote over three open-source VLMs. Those answers are used to calculate three metrics: *action accuracy*, *target accuracy*, and *stillness accuracy*. These metrics measure, respectively, whether the actor and action match, whether it is directed at the correct target, and whether non-acting persons remain still.

We apply SocialDirector to two open-source generative models: Wan2.2 [62] and LTX2.3 [19]. On both backbones, our method consistently improves social interaction fidelity over the uncontrolled baselines, with gains of up to 3.9%, 3.1%, and 3.1% on the three metrics. Notably, these results closely match the performance of a GT-oracle variant equipped with ground-truth annotations, suggesting that our generated interactions approach the upper bound set by real videos. Our main contributions are summarized as follows:

- We propose SocialDirector, a training-free interaction controller that enhances the video diffusion model for multi-person video generation. It controls *who* performs *which* action, *when* it occurs, and *toward whom* it is directed through modulating cross-attention maps.
- We annotate existing datasets and build the first fully automated evaluation pipeline for social interaction video generation. It is powered by majority vote over open-source VLMs.

- Evaluations demonstrate that SocialDirector significantly improves social interaction fidelity and achieves state-of-the-art performance against competitive baselines.

2 Related Works

2.1 Multimodal Social Interaction

Human social interaction is inherently multimodal, spanning spoken language [25], facial expressions [22], gaze [89, 40, 37], gestures [5, 38], and body movements [3, 39]. The community has developed diverse tasks and benchmarks to study these phenomena, including video question answering [82, 28], conversational modeling [59, 32, 23], speaker prediction [47, 46], and behavior classification [5, 51], covering scenarios such as board games [31, 87], daily conversations [34], and meetings [44]. Recent work has increasingly adopted multimodal foundation models for social interaction understanding [43, 33, 50, 26]. On the generation side, several methods synthesize multi-person 3D motion sequences conditioned on interaction categories or partner movements [79, 36, 75]. Compared with these understanding- and motion-level efforts, our work is the first to tackle *controllable video generation* of multimodal social interactions.

2.2 Multi-Person Video Generation

Human-centric video generation has been extensively studied, spanning talking video synthesis [65, 8], portrait animation [17, 11], and full-body motion generation [24, 12, 53]. Extending to multi-person scenarios [29, 13, 41, 21, 88, 91, 90, 64, 14, 25, 7] has recently attracted growing attention, where the central challenge has been correctly binding each person's audio to the corresponding individual in the scene. Existing approaches address this through label rotary position embedding [29], 3D-mask-based embedding routers [21], identity-aware audio-face cross attention [88], and mask-guided classifier-free guidance [41]. Unlike these audio-driven methods, our work focuses on the social interaction behaviors of multiple persons in generated videos—including action types, facial expressions, and especially interaction targets (i.e., who interacts with whom)—which have been largely overlooked by prior work. Furthermore, our approach is entirely training-free through cross-attention manipulation at inference time, whereas most existing multi-person video generation methods require dedicated multi-person training data and task-specific fine-tuning.

2.3 Cross-Attention Control in Diffusion Models

In diffusion models, cross-attention serves as the key mechanism for conditioning generation on external signals such as text prompts and reference images, applicable across both U-Net [58, 57] and Diffusion Transformer [52] backbones. Cross-attention control has been widely adopted for text-to-image generation [27, 6, 56], image editing [20, 63, 1], layout-guided generation [42, 9, 73], subject-driven synthesis [68, 66, 72], multi-subject binding [15], noise optimization [18, 49], and identity preservation [86, 74, 81]. Constraining attention with spatial or temporal windows and attention masks is also a common strategy for controlling generation results [15, 73, 84]; for example, Bounded Attention [15] bounds both self- and cross-attention of each subject, and TS-Attn [84] applies a soft temporal bias for multi-event video generation. Related attention modulation has also been explored for video editing and compositional video generation [78, 70, 83]. Compared with these methods, we are the first to bring such control to multi-person social interaction, and the control over the *interaction target*, i.e., toward whom an action is directed, has not been addressed by prior work.

3 Proposed Method

We propose **SocialDirector**, a training-free controller that enhances multi-person video generation with explicit control over social interactions. It consists of two modules that modulate the cross-attention map of each DiT block: *Social Actor Masking* (Section 3.2), which applies a spatial-temporal mask that binds each social event to its own actor, and *Directional Reweighting* (Section 3.3), which guides actions toward correct targets. An overview is illustrated in Figure 2.

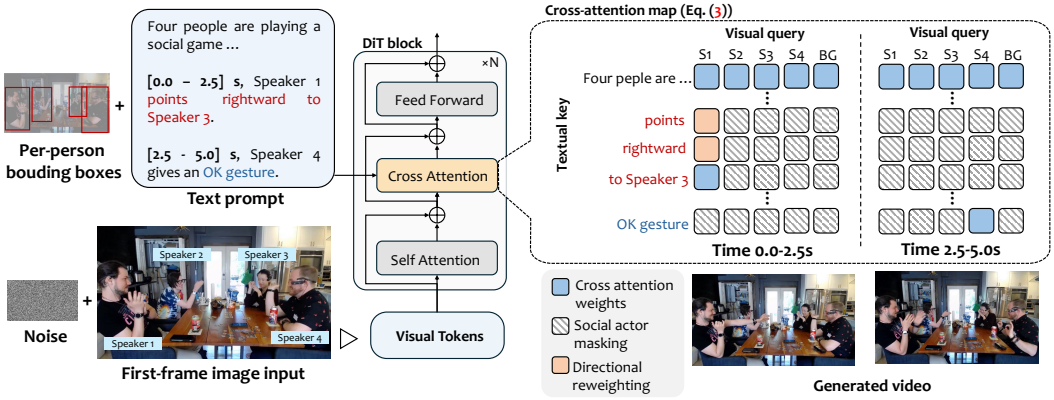


Figure 2: Overview of the proposed method. Given a first-frame image, per-person bounding boxes, and a structured text prompt describing per-person social events, our method generates multi-person videos with faithful social interactions. The generation is controlled through two modules modulating the cross-attention map: *Social Actor Masking* and *Directional Reweighting*.

3.1 Multimodal Social Interaction Video Generation

As shown in Figure 2, given an RGB first-frame image I_0 containing N persons, a set of per-person bounding boxes $\mathcal{B} = \{b_s \mid s \in [1, N]\}$ of the image, and a text prompt $\mathcal{P}$, our social interaction video generation model outputs a video faithful to $\mathcal{P}$. Different from general multi-person generation, $\mathcal{P}$ is organized as a shared scene description followed by a sequence of K *structured social events*, each of the form "$[t_1^k, t_2^k]$, person s performs action a (toward person s')". The $[t_1^k, t_2^k]$ is the time window of the k-th event, $s \in [1, N]$ is the actor, a is the action, and s' is an optional interaction target. A faithful generation requires joint control over *who* acts, *what* action is performed, *when* it occurs, and *toward whom* it is directed.

Attention leakage in standard DiTs. In a standard image-to-video DiT pipeline, I_0 and the added noise are tokenized into visual tokens $\mathcal{V} = \{v_i \in \mathbb{R}^d \mid i \in [1, T \times H \times W]\}$, and $\mathcal{P}$ is encoded into text tokens $\mathcal{L} = \{l_j \in \mathbb{R}^d\}$. Stacking these tokens as rows yields matrices V and L, which serve respectively as queries and keys in the cross-attention of each DiT block:

$$\text{Attn} = \text{softmax}_l\left(\frac{VL^\top}{\sqrt{d}}\right). \tag{1}$$

This equation is agnostic to person identity and event time: a visual token belonging to one person can freely attend to text describing another person, at any moment in the video. As a result, actions are often performed by the wrong person, outside their annotated time window, or directed at the wrong target. We refer such failure as *attention leakage*. Attention leakage has been extensively studied in text-to-image generation, where cross-attention control is widely adopted for manipulating layout, attribute binding, and semantic composition in multi-subject outputs [63, 6, 56].

3.2 Social Actor Masking

To address attention leakage in multi-person video generation, we propose *Social Actor Masking*, which binds each social event to its own actor in both space and time through a cross-attention mask. We encode $\mathcal{P}$ into a background token $\mathcal{L}_{\text{bg}}$ and per-event tokens $\mathcal{L}_{s,k}$. The background token $\mathcal{L}_{\text{bg}}$ describes the shared scene, while the per-event token $\mathcal{L}_{s,k}$ contains the action text of the k-th event associated with person s. For visual tokens, we use the per-person bounding box b_s and the event time window $[t_1^k, t_2^k]$ to localize the actor-specific spatiotemporal subset $\mathcal{V}_{s,k} \subset \mathcal{V}$ for each event. We then define a mask M as:

$$M_{v,l} = \begin{cases} 0, & l \in \mathcal{L}_{\text{bg}}, \\ 0, & l \in \mathcal{L}_{s,k} \text{ and } v \in \mathcal{V}_{s,k}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

As illustrated by the cross-attention map of Figure 2, the background span remains accessible to every visual query while each event span is gated to the visual tokens of its own actor within the relevant

time window. The spatial dimension of $\mathcal{V}_{s,k}$ enforces *who does what*, and its temporal dimension enforces *when*, eliminating cross-person and cross-time attention leakage.

3.3 Directional Reweighting

Section 3.2 has no constraint on the action's target. As a result, mistargeted actions remain even when actor and time are correctly bound. To address this, we introduce a *Directional Reweighting* module. Specifically, we first derive a direction (e.g., "leftward" or "rightward") from the actor's and target's bounding boxes, and insert the corresponding directional phrase (e.g., "points rightward" in Figure 2) into the text prompt. We denote the set of all such directional tokens by $\mathcal{L}_{\text{dir}}$. We then introduce a positive bias matrix W that amplifies attention to these tokens, as illustrated by the cross-attention map of Figure 2. Together with M in Equation (2), W is added to the cross-attention map:

$$
\begin{aligned}
W_{v,l} &= \begin{cases} \gamma \cdot \sqrt{d}, & l \in \mathcal{L}_{s,k} \cap \mathcal{L}_{\text{dir}} \text{ and } v \in \mathcal{V}_{s,k}, \\ 0, & \text{otherwise}, \end{cases} \\
\widehat{\text{Attn}} &= \text{softmax}_l\left(\frac{VL^\top}{\sqrt{d}} + M + W \right),
\end{aligned}
\tag{3}
$$

where the reweighting stength γ is a pre-defined hyper-parameter. Scaling the bias by $\sqrt{d}$ matches the intrinsic scale of the unnormalized logits $VL^\top$, so that the γ remains decoupled from the hidden dimension. With $\gamma > 0$, this module sharpens the model's sensitivity to directional cues, guiding each action toward its intended target.

Both Social Actor Masking and Directional Reweighting operate purely on the cross-attention map at inference, introducing no additional parameters. SocialDirector is therefore plug-and-play for any DiT-based image-to-video backbone, requiring no architectural modification or fine-tuning.

4 Evaluation Pipeline

We propose a fully automated evaluation pipeline for social interaction video generation. Existing metrics are designed for video quality and identity consistency, failing to capture multi-person interaction fidelity. Recent works [91, 88] on multi-person video evaluation focus on audio-visual synchronization rather than social interactions, e.g., verifying the correctness of *who* performs *which* action toward *whom*. To fill this gap, we construct an evaluation dataset from existing social interaction benchmarks and employ open-source VLMs to assess generated videos. An overview of the pipeline is provided in Figure 3.

4.1 Evaluation Dataset

We collect ground-truth video clips from three publicly available multi-person interaction datasets, selecting clips with rich social interactions via the per-dataset procedures described below:

MELD [54] is a multimodal multi-party dataset for emotion recognition in conversations, with annotations focused on emotional expressions during social exchanges. We select 19 five-second clips in which the scene and visible persons remain unchanged throughout.

MMSI [32] is a social interaction dataset with annotations of speaking behavior modeling including speaking targets. We select 50 clips that exhibit rich social dynamics (three or more speaking-target events within five seconds), retaining the original speaking-target and timestamp annotations.

SocialGesture [5] is a multi-person gesture recognition dataset captured from social game recordings. We randomly sample 80 annotated clips and ask annotators to re-annotate temporal timestamps within five seconds and additional social behaviors on top of the original gesture and target labels.

In total, the evaluation dataset contains 149 multi-person video clips, 674 annotated speakers, and 479 action events (299 with a specific interaction target), spanning 11 coarse action categories such as speaking, object interaction, and gestures. We additionally curate first-frame bounding-box annotations for every person, with manual corrections to ensure spatial accuracy.

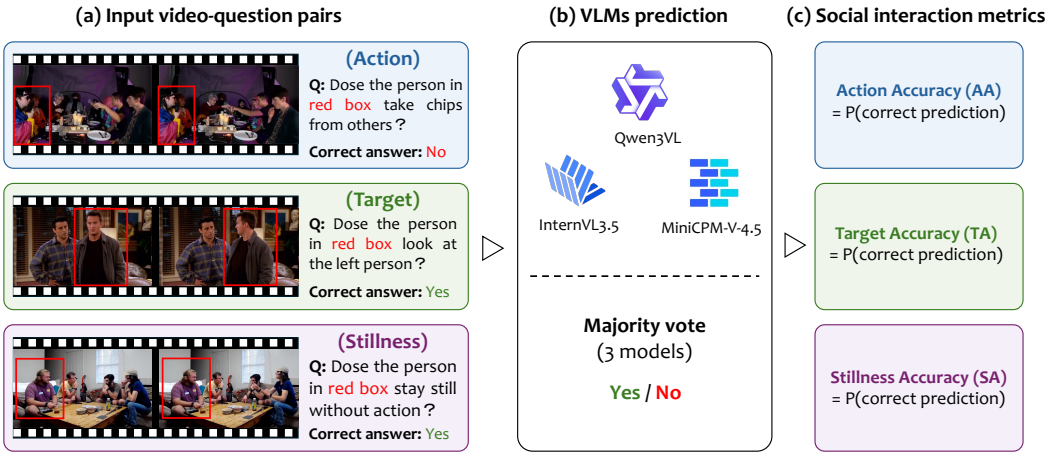


Figure 3: Overview of our evaluation pipeline. (a) Each annotated social event is converted into a video-question pair. (b) Three VLMs independently answer each question (yes or no) and their predictions are aggregated via majority voting. (c) The voting results, along with the correct answer, yield three social interaction metrics: Action Accuracy, Target Accuracy, and Stillness Accuracy.

4.2 Social Interaction Metrics with VLMs

A faithful social interaction video should correctly realize who performs what action, when it occurs, and toward whom it is directed. To evaluate these aspects, we design three complementary metrics targeting different dimensions of social interaction fidelity. **Action accuracy** verifies whether the designated person performs the correct action category within the annotated time window. For each annotated event, we generate a *positive* query asking whether the correct person performs the expected action, and a *negative* query asking whether another person performs that action, with expected answers "yes" and "no" respectively. **Target accuracy** further assesses whether an action is toward the correct target, through a similar positive/negative query pair that substitutes a wrong target. This metric is evaluated only on events with an explicit interaction target. **Stillness accuracy** measures whether persons with no annotated events correctly remain still, computed only on uninvolved speakers. Together, action accuracy evaluates "who does what, when", target accuracy evaluates "toward whom", and stillness accuracy penalizes actions that are not specified, providing a multi-dimensional social interaction evaluation.

We formulate the above metrics binary visual question answering and evaluate each with three open-source VLMs. As illustrated in Figure 3, each question is posed on a temporally trimmed sub-clip centered on the annotated event, so that the VLMs observe only the relevant time window. To specify the correct actor in multi-person scenes, we overlay a colored bounding box around the person of interest on the video [35, 77, 60]. In total, the question set comprises 479 positive and 399 negative queries for action accuracy, 292 positive and 263 negative queries for target accuracy, and 314 queries for stillness accuracy. We employ three VLMs (Qwen3-VL [2], InternVL3.5 [67], MiniCPM-V 4.5 [80]) and aggregate their answers via majority voting. More details about our evaluation metrics can be found in Section A.2. We further validate the reliability of the VLM judges against human annotators in Section A.3.

4.3 Video Quality Metrics

Beyond social interaction metrics, we adopt five complementary video quality metrics following previous multi-person video generation works [29, 88]: FVD [61] for distributional similarity to ground-truth videos, LPIPS [85] for frame-level perceptual distance, ViCLIP [69] similarity for text prompt alignment, DOVER [71] for aesthetic video quality, and DINOv2 [48] cosine similarity for per-person identity preservation across frames [86, 81, 74].

Table 1: Comparison of SocialDirector with baseline methods on our social interaction benchmark. Green indicates improvement over the base model; red indicates degradation.

Method	Box Input	Social Interaction Metrics			Video Quality Metrics				
		Action Acc ↑	Target Acc ↑	Stillness Acc ↑	FVD ↓	LPIPS ↓	ViCLIP ↑	Vid Quality ↑	ID Pres ↑
GT Oracle		75.9	72.6	90.8	—	—	15.6	39.0	88.8
EchoMimic v3		60.1	65.5	**98.0**	5.26	17.6	14.4	35.9	90.8
AnyTalker		64.1	66.3	92.9	5.40	16.8	15.2	36.3	**91.7**
Bind-Your-Avatar	✓	63.6	69.5	72.2	20.1	31.2	16.4	28.4	66.9
MultiTalk	✓	68.1	68.6	94.1	11.2	22.3	15.0	41.7	89.9
HunyuanVideoAvatar		68.9	68.6	88.5	5.06	18.8	16.5	36.2	86.8
LTX-2.3		71.6	69.4	87.6	5.77	17.7	15.9	31.8	87.0
LTX-2.3 + SocialDirector	✓	74.3 +2.7	70.8 +1.4	88.3 +0.7	5.89 +0.12	17.7 0.0	16.4 +0.5	32.4 +0.6	86.7 −0.3
Wan2.2		72.2	69.2	85.1	**4.21**	17.2	16.1	43.3	84.1
Wan2.2 + SocialDirector	✓	**76.1** +3.9	**72.3** +3.1	88.2 +3.1	4.29 +0.08	**16.2** −1.0	**16.8** +0.7	**43.7** +0.4	85.1 +1.0

5 Experiments

5.1 Baseline Methods

We apply SocialDirector on top of two recent image-to-video DiTs: Wan2.2 [62] and LTX2.3 [19]. We further compare against competitive multi-person video generation methods: MultiTalk [29], EchoMimic v3 [45], HunyuanVideoAvatar [10], Bind-Your-Avatar [21], AnyTalker [88], and Playmate2 [41]. For fair comparison, all methods receive the same first-frame image and text prompt, where the text prompt explicitly describes the relative position and action of each speaker. Methods that additionally accept per-person bounding-box inputs [29, 41, 21] are given identical boxes. Since several of the baselines require an audio input, we supply a silent audio track to remove its impact. More details about baseline methods are provided in Section A.1.

5.2 Implementation Details

Our main experiments use the pretrained Wan2.2-I2V-A14B [62] and LTX-2.3-22B [19] checkpoints. We generate videos at 832×480 resolution with 81 frames using both Wan2.2 and LTX-2.3. We use 40 denoising steps for Wan2.2 and 30 for LTX-2.3, following the officially recommended settings of each model. Our method is applied to all cross-attention layers of every DiT block. For Social Actor Masking, we empirically expand each person's bounding box by 15% on each side to accommodate natural body movement beyond the first-frame annotation. For Directional Reweighting, we apply it to directional words and their verbs, e.g., "pointing leftward" and "speaks to right". The reweighting strength is set to $\gamma = 0.5$. All methods are run three times with the same set of random seeds, and we report the averaged results. Each video is generated on a single NVIDIA L40S GPU.

5.3 Results

Comparison with baselines. As shown in Table 1, applying SocialDirector to both Wan2.2 and LTX-2.3 yields consistent improvements on all three social interaction metrics. On Wan2.2, action accuracy improves by 3.9%, target accuracy by 3.1%, and stillness accuracy by 3.1% over the base model, demonstrating the effectiveness of our method in resolving attention leakage and strengthening action-target binding. Compared with other multi-person video generation methods, SocialDirector achieves state-of-the-art performance on action and target accuracy. In particular, SocialDirector already matches the GT oracle on action and target accuracy on Wan2.2, reaching the upper bound set by real videos. Per-VLM (Table 8) and per-seed (Table 9) analyses in the Appendix show that these gains hold across individual VLMs and random seeds, confirming that they are not artifacts of any particular random seed. Meanwhile, the video quality metrics of SocialDirector are comparable to the base Wan2.2 and LTX-2.3, indicating that our method preserves visual quality.

For stillness accuracy and identity preservation, several audio-driven methods [45, 88] exhibit abnormally high scores that even exceed the GT oracle. This occurs because these methods are trained primarily to align with audio streams and overlook text-driven controls. Their outputs tend to produce no action, which trivially inflates both metrics. Other existing methods [29, 45, 10, 21] perform poorly on social interaction metrics, as their implementations do not support scenes with three or more persons. A robustness analysis under noisy boxes and timestamps and a comparison with the commercial model Kling 3.0 [30] are provided in Sections A.4 and A.5.

Table 2: Comparison with training-free attention control methods on Wan2.2.

Method	Social Interaction Metrics			Video Quality Metrics		Cost	
	Action Acc ↑	Target Acc ↑	Stillness Acc ↑	FVD ↓	ID Pres ↑	Time ↓	Peak VRAM ↓
Wan2.2	72.2	69.2	85.1	**4.21**	84.1	1.00×	**34.5 GB**
+ Bounded Attention [15]	70.4	69.5	**88.9**	7.38	70.1	2.19×	37.0 GB
+ BoxDiff [73]	63.8	69.0	80.9	19.98	49.0	1.39×	62.2 GB
+ TS-Attn [84]	75.9	68.6	86.5	4.74	82.2	1.05×	38.8 GB
+ SocialDirector (ours)	**76.1**	**72.3**	88.2	4.29	**85.1**	1.03×	**34.5 GB**



Figure 4: Qualitative comparison with baseline methods on multi-person social interaction generation. All methods use the same random seed.

Comparison with training-free attention control methods. We further compare SocialDirector with Bounded Attention [15], BoxDiff [73], and TS-Attn [84], all applied to the same Wan2.2 backbone with identical inputs. As shown in Table 2, SocialDirector achieves the best action accuracy, target accuracy, and identity preservation, and is the only method that improves target accuracy over the base model, while its FVD stays closest to the base model. It is also the most efficient: Bounded Attention more than doubles the generation time and BoxDiff nearly doubles the peak memory, whereas SocialDirector adds only 3% time and no extra memory. These results show that our method is both simple and effective for social interaction control.

Visualizations. Qualitative comparisons in Figure 4 highlight the advantages of SocialDirector over existing baselines. AnyTalker [88] remains largely insensitive to action controls from the text modality, resulting in nearly static videos. Bind-Your-Avatar [21] suffers from significant degradation in visual quality when scaled from its two-person implementation to complex 4–6 person scenarios. Although MultiTalk [29] and base Wan2.2 models generate the specified actions, they assign them to a wrong actor or direct them toward a wrong target, confirming the existence of attention leakage. In contrast, SocialDirector consistently generates faithful social interactions, precisely following the "who," "what," and "to whom" from text prompts. Qualitative comparisons against more methods are provided in Section A.6.

Table 3: Ablation on SocialDirector's components based on Wan2.2. M: Social Actor Masking, W: Directional Reweighting.

M	W	Action Acc ↑	Target Acc ↑	Stillness Acc ↑
		72.2	69.2	85.1
✓		75.7	69.7	87.3
	✓	73.0	68.6	84.1
✓	✓	**76.1**	**72.3**	**88.2**

Table 4: Per-sample computational cost of SocialDirector.

Method	Time (s)	FLOPs ($\times 10^{15}$)
Wan2.2	604.0	37.53
Wan2.2 + Ours	614.2 +1.7%	37.53
LTX-2.3	180.4	33.79
LTX-2.3 + Ours	185.2 +2.7%	33.79



Figure 5: Qualitative ablations of SocialDirector.

5.4 Ablations

Effect of Attention Control Components. We ablate Social Actor Masking and Directional Reweighting separately, i.e., M and W in Equation (3). As shown in Table 3, applying M alone substantially improves action and stillness accuracy, confirming that attention leakage has been mitigated. However, gains in target accuracy remain limited. Adding W further provides directional cues from the text side, boosting target accuracy by a large margin. In contrast, only applying W offers no improvement, as it leaves attention leakage unresolved.

As visualized in Figure 5 (a), W guides actions toward intended targets. Specifically, when using only M, the model lacks awareness of target directions, resulting in an incorrect video where all speakers perform a collective toasting in the center. With $M + W$, Speaker 3 and Speaker 4 correctly toast to the left and right respectively, forming the two distinct interacting groups specified in the prompt. These two modules together achieve effective social interaction control.

Effect of Temporal Masking. We examine the effect of enabling the temporal dimension of Social Actor Masking. As shown in Figure 5 (b), disabling temporal masking leads to a disorder in the sequence of events, where actions fail to respect their scheduled timings. Enabling temporal masking confines each action to its annotated time window and restores the correct event order, ensuring that the generated social dynamics faithfully follow the specified interaction timeline.

Computation Cost. Table 4 reports the inference overhead introduced by SocialDirector. On Wan2.2, the generation time increases by only 1.7%, and on LTX-2.3 by 2.7%. The additional FLOPs are negligible, as our method only performs element-wise additions on the cross-attention score matrices without introducing any new parameters or forward passes. This confirms the efficiency of our training-free approach.

Table 5: Performance of different reweighting strength γ. All experiments are based on Wan2.2.

γ	Social Interaction Metrics			Video Quality Metrics				
	Action Acc ↑	Target Acc ↑	Stillness Acc ↑	FVD ↓	LPIPS ↓	ViCLIP ↑	Vid Quality ↑	ID Pres ↑
0	75.7	69.7	87.3	**4.11**	16.4	16.4	43.6	**85.8**
0.1	75.8	70.6	86.9	4.15	16.5	16.5	43.6	85.7
0.3	**76.2**	71.6	87.6	4.15	16.3	16.1	43.6	85.6
0.5	76.1	**72.3**	88.2	4.29	**16.2**	**16.8**	**43.7**	85.1
0.7	75.6	69.4	88.5	4.84	17.2	16.2	41.7	83.2
1.0	72.6	67.9	**90.9**	6.55	21.0	15.3	41.2	79.6

Reweighting Strength. Table 5 reports the sensitivity of our method to the reweighting strength γ. Smaller values provide insufficient directional bias, leading to lower target accuracy, whereas larger values over-amplify the directional tokens and degrade video quality. We select $\gamma = 0.5$, which achieves the best overall performance.

5.5 User Study

To complement the automatic metrics, we conduct a pairwise user study with 17 participants. In each trial, a participant is shown the interaction prompt and two videos generated from the same inputs, one by SocialDirector and one by a competing method (the base model, MultiTalk, or HunyuanVideoAvatar), in randomized order, and asked (i) which video better matches the described interaction, and (ii) in which video the interaction looks more natural. Table 6 reports the percentage of judgments preferring SocialDirector. Our videos are clearly preferred on description match against all three competitors (78.0–88.3%). On naturalness, SocialDirector is strongly preferred over HunyuanVideoAvatar (89.5%) and on par with the base model (54.0%, with the confidence interval covering 50%). This is expected: our training-free attention control does not alter the backbone's ability to generate natural motion, but only controls who performs which action toward whom and when.

Table 6: User study. Each cell reports the percentage of judgments preferring SocialDirector over the competing method, with the bootstrap 95% confidence interval over raters. N is the number of judgments.

SocialDirector *vs.*	N	Description Match	Action Naturalness
Base (same backbone)	122	78.0 [72.1, 84.2]	54.0 [43.9, 66.1]
MultiTalk	118	88.3 [80.4, 95.3]	58.6 [50.0, 67.9]
HunyuanVideoAvatar	114	82.5 [73.7, 90.7]	89.5 [80.7, 96.5]

6 Conclusion

This paper presents **SocialDirector**, a training-free controller that enhances pretrained image-to-video DiTs to faithfully generate multi-person social interaction videos. The proposed Social Actor Masking strengthens the binding between visual and textual tokens of the same actor across space and time. The proposed Directional Reweighting amplifies the attention to directional tokens, guiding each action toward its intended target. We further contribute a benchmark and the first fully automated VLM-based evaluation pipeline. Extensive experiments demonstrate that SocialDirector substantially improves interaction fidelity without compromising visual quality.

Limitations and Future Directions. Our current work has several limitations that point to natural directions for future research. First, SocialDirector operates in the visual-text domain and does not yet integrate the audio modality, falling short on speech-grounded behaviors such as lip synchronization. Second, our method is evaluated on short clips and does not support long videos with sustained, multi-turn social dynamics. Third, precise control over actor, target, and timing requires *structured inputs*, i.e., a prompt with per-event timestamps and per-person bounding boxes; free-form instructions must first be parsed into this format, e.g., by an LLM. Fourth, when events overlap in space or time, their visual tokens attend to the union of the corresponding text spans, so attention leakage may remain; under heavy overlap, severe occlusion, or large actor movements, a static box-based mask can no longer separate the actors and our control degrades, although moderate noise in boxes and timestamps is well tolerated (Section A.4). Finally, VLM-based evaluation may misjudge subtle social behaviors, which we mitigate by majority voting and human validation (Section A.3). Building on these limitations, future research includes incorporating audio for verbal and non-verbal signals, coupling with streaming avatar models for long-form generation, replacing static box-based attention bias with dynamic, tracking-based conditioning, and replacing manual interaction scripts with an LLM-based social planner that decides who should interact with whom and when, moving video synthesis closer to autonomous, socially intelligent agents.

Acknowledgments and Disclosure of Funding

This work was supported by JST ASPIRE (Grant Number JPMJAP2303), JSPS Grant-in-Aid for Scientific Research (B) (Grant Number JP24K02956), and Grant-in-Aid for JSPS Fellows (Project Number JP25KJ1180). The authors declare no competing interests.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 conference papers*, pages 1–12, 2024.

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

[3] Michal Balazia, Philipp Müller, Ákos Levente Tánczos, August von Liechtenstein, and François Brémond. Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 70–79, 2022.

[4] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

[5] Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19509–19519, 2025.

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[7] Junhao Chen, Mingjin Chen, Jianjin Xu, Xiang Li, Junting Dong, Mingze Sun, Puhua Jiang, Hongxiang Li, Yuhang Yang, Hao Zhao, et al. Dancetogether! identity-preserving multi-person interactive video generation. *arXiv preprint arXiv:2505.18078*, 2025.

[8] Ming Chen, Liyuan Cui, Wenyuan Zhang, Haoxian Zhang, Yan Zhou, Xiaohan Li, Songlin Tang, Jiwen Liu, Borui Liao, Hejia Chen, et al. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation. *arXiv preprint arXiv:2508.19320*, 2025.

[9] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5343–5353, 2024.

[10] 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.

[11] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2403–2410, 2025.

[12] 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.

[13] Ernie Chu and Vishal M Patel. Face-to-face: A video dataset for multi-person interaction modeling. *arXiv preprint arXiv:2603.14794*, 2026.

[14] Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. Unils: End-to-end audio-driven avatars for unified listening and speaking. *arXiv preprint arXiv:2512.09327*, 2025.

[15] Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision (ECCV)*, 2024.

[16] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[17] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

[18] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9380–9389, 2024.

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

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[22] Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. Smile: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1149–1167, 2024.

[23] Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M Rehg, Vamsi Krishna Ithapu, and Ruohan Gao. The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26396–26405, 2024.

[24] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22747–22757, 2023.

[25] Zeyu Jin, Songtao Zhou, Haoyu Wang, Minghao Tian, Kaifeng Yun, Zhuo Chen, Xiaoyu Qin, and Jia Jia. From natural alignment to conditional controllability in multimodal dialogue. In *The Fourteenth International Conference on Learning Representations*, 2026.

[26] Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, Ruicong Liu, and Yoichi Sato. Can mllms read the room? a multimodal benchmark for assessing deception in multi-party social interactions. *arXiv preprint arXiv:2511.16221*, 2025.

[27] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023.

[28] Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Sivbench: A video benchmark for social interaction understanding and reasoning. *arXiv preprint arXiv:2506.05425*, 2025.

[29] 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[30] Kuaishou Technology. Kling ai 3.0. https://klingai.com, 2026.

[31] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6570–6588, 2023.

[32] Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14595, 2024.

[33] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024.

[34] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379, 2018.

[35] Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M Rehg, and Yapeng Tian. Towards online multi-modal social interaction understanding. *arXiv preprint arXiv:2503.19851*, 2025.

[36] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024.

[37] Ruicong Liu, Yunfei Liu, Haofei Wang, and Feng Lu. Pnp-ga+: Plug-and-play domain adaptation for gaze estimation using model variants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3707–3721, 2024.

[38] Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2024.

[39] Ruicong Liu, Yifei Huang, Liangyang Ouyang, Caixin Kang, and Yoichi Sato. Sfhand: A streaming framework for language-guided 3d hand forecasting and embodied manipulation. *arXiv preprint arXiv:2511.18127*, 2025.

[40] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3835–3844, 2021.

[41] Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7818–7826, 2026.

[42] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5321–5329, 2023.

[43] Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in ai agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20541–20560, 2024.

[44] Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, pages 137–140, 2005.

[45] Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8008–8015, 2026.

[46] Philipp Müller, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Michael Dietz, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4878–4882, 2021.

[47] Curtis G Northcutt, Shengxin Zha, Steven Lovegrove, and Richard Newcombe. Egocom: A multi-person multi-modal egocentric communications dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6783–6793, 2023.

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[49] Liangyang Ouyang and Jiafeng Mao. Lore: Latent optimization for precise semantic control in rectified flow-based image editing. *arXiv preprint arXiv:2508.03144*, 2025.

[50] Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. Multi-speaker attention alignment for multimodal social interaction. *arXiv preprint arXiv:2511.17952*, 2025.

[51] Liangyang Ouyang, Yuki Sakai, Ryosuke Furuta, Hisataka Nozawa, Hikoro Matsui, and Yoichi Sato. Leadership assessment in pediatric intensive care unit team training. *arXiv preprint arXiv:2505.24389*, 2025.

[52] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[53] Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, et al. Actavatar: Temporally-aware precise action control for talking avatars. *arXiv preprint arXiv:2512.19546*, 2025.

[54] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536, 2019.

[55] Di Qiu, Zheng Chen, Rui Wang, Mingyuan Fan, Changqian Yu, Junshi Huang, and Xiang Wen. Moviecharacter: A tuning-free framework for controllable character video synthesis. *arXiv preprint arXiv:2410.20974*, 2024.

[56] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.

[57] 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.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[59] Fiona Ryan, Hao Jiang, Abhinav Shukla, James M Rehg, and Vamsi Krishna Ithapu. Egocentric auditory attention localization in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14663–14674, 2023.

[60] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023.

[61] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[62] 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.

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

[64] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025.

[65] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.

[66] Weicheng Wang, Guoli Jia, Zhongqi Zhang, Liang Lin, and Jufeng Yang. Ps-diffusion: Photorealistic subject-driven image editing with disentangled control and attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18302–18312. IEEE Computer Society, 2025.

[67] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[68] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *The Thirteenth International Conference on Learning Representations*, 2025.

[69] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[70] Zun Wang, Jialu Li, Han Lin, Jaehong Yoon, and Mohit Bansal. Dreamrunner: Fine-grained compositional story-to-video generation with retrieval-augmented motion adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[71] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20144–20154, 2023.

[72] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

[73] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7452–7461, 2023.

[74] Qi Xie, Yongjia Ma, Donglin Di, Xuehao Gao, and Xun Yang. Moca: Identity-preserving text-to-video generation via mixture of cross attention. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pages 1–8, 2025.

[75] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22260–22271, 2024.

[76] Zhenran Xu, Longyue Wang, Jifang Wang, Zhouyi Li, Senbao Shi, Xue Yang, Yiyu Wang, Baotian Hu, Jun Yu, and Min Zhang. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909*, 2025.

[77] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006, 2023.

[78] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *International Conference on Learning Representations (ICLR)*, 2025.

[79] Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. Socialgen: Modeling multi-human social interaction with language models. In *Thirteenth International Conference on 3D Vision*, 2026.

[80] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

[81] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12978–12988, 2025.

[82] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[83] Hongyu Zhang, Yufan Deng, Shenghai Yuan, Yian Zhao, Peng Jin, Xuehan Hou, Chang Liu, and Jie Chen. Comp-attn: Present-and-align attention for compositional video generation. *arXiv preprint arXiv:2503.14428*, 2025.

[84] Hongyu Zhang, Yufan Deng, Zilin Pan, Peng-Tao Jiang, Bo Li, Qibin Hou, Zhiyang Dou, Zhen Dong, and Daquan Zhou. Ts-attn: Temporal-wise separable attention for multi-event video generation. In *International Conference on Learning Representations (ICLR)*, 2026.

[85] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[86] Yuechen Zhang, Yaoyang Liu, Bin Xia, Bohao Peng, Zexin Yan, Eric Lo, and Jiaya Jia. Magicmirror: Id-preserved video generation in video diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14464–14474, 2025.

[87] Zheng Zhang, Nuoqian Xiao, Qi Chai, Deheng Ye, and Hao Wang. Multimind: Enhancing werewolf agents with multimodal reasoning and theory of mind. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5824–5833, 2025.

[88] Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

[89] Qi Zhou, Wannapon Suraworachet, and Mutlu Cukurova. Detecting non-verbal speech and gaze behaviours with multimodal data and computer vision to interpret effective collaborative learning interactions. *Education and information technologies*, 29(1):1071–1098, 2024.

[90] Yang-Hao Zhou, Haitian Li, Rexar Lin, Heyan Huang, Jinxing Zhou, Changsen Yuan, Tian Lan, Ziqin Zhou, Yudong Li, Jiajun Xu, et al. Mtavg-bench: A comprehensive benchmark for evaluating multi-talker dialogue-centric audio-video generation. *arXiv preprint arXiv:2602.00607*, 2026.

[91] Yingjie Zhou, Xilei Zhu, Siyu Ren, Ziyi Zhao, Ziwen Wang, Farong Wen, Yu Zhou, Jiezhang Cao, Xiongkuo Min, Fengjiao Chen, et al. Evaltalker: Learning to evaluate real-portrait-driven multi-subject talking humans. *arXiv preprint arXiv:2512.01340*, 2025.

A Technical Appendix

A.1 Baseline Details

This section provides implementation details for each baseline method used in our experiments. All baseline methods receive the same first-frame image and text prompt as SocialDirector. For methods that require audio input, we supply a 5s silent waveform so that the comparison isolates each model's ability to generate social interaction behaviors from visual and textual cues alone. The model-specific configurations are detailed below.

- **Wan2.2** [62] is a state-of-the-art open-source video generation model. We use the I2V-A14B checkpoint (Wan-AI/Wan2.2-I2V-A14B) as our main base model.
- **LTX-2.3** [19] is another leading open-source video generation model. We use the 22B checkpoint (Lightricks/LTX-2.3) as a secondary base model to verify generalizability.
- **MultiTalk** [29] is built on Wan2.1-I2V-14B and injects per-speaker audio embeddings into the diffusion backbone. Its inference pipeline natively supports at most two speakers; for scenes with three or more persons, we select the two persons whose action events occur earliest and drop the rest.
- **Playmate2** [41] is built on Wan2.1-I2V-14B-720P and accepts an arbitrary speaker count. Because this method requires separated per-speaker audio inputs, which are unavailable in our multi-person dataset, and its generation speed is prohibitively slow under our setting and hardware, we only include it in qualitative comparisons; for these qualitative cases, we split the native audio track into temporal segments and assign them to 4–6 speakers.
- **AnyTalker** [88] is built on the 1.3B Wan2.1-Fun-V1.1-InP backbone and dynamically matches conditioned faces to the input audio list. We feed our annotated per-person bounding boxes together with silent audio tracks into its Audio-Face Attention Mask, so that generation is driven purely by the first frame and text prompt.
- **Bind-Your-Avatar** [21] is a CogVideoX-based method that requires exactly two cropped face references by InsightFace. We select the two earliest-event persons in the same way as MultiTalk. Because the model is constrained to 49 frames at 25,fps by CogVideoX's positional embedding, we temporally stretch the output to approximately 5s to match the duration of other baselines.
- **HunyuanVideo-Avatar** [10] features a Face-Aware Audio Adapter but its released inference script processes a single image–audio–prompt tuple and only detects the largest face per frame, effectively operating as a single-person baseline on our multi-person scenes.
- **EchoMimic v3** [45] is a single-subject audio-driven portrait animation method based on Wan2.1-Fun-1.3B, included as a high-quality single-person baseline.

A.2 Evaluation Details

This section provides additional details on the evaluation pipeline introduced in Section 4, including the generation prompt format, action category taxonomy, and VLM query design.

Generation Prompt Format. To generate videos for evaluation, we construct a structured text prompt for each sample. The prompt begins with a scene-level description enumerating all speakers by their left-to-right position, followed by one sentence per speaker describing their action. Below are two representative examples with *directional words* emphasized:

- "There are 3 people in the scene: the person on the left (speaker 1), the person in the middle (speaker 2), the person on the right (speaker 3). [0s–3s] The person on the left listening while touching his chin. [1s–4s] The person in the middle *speaks leftward* to speaker 1 with anger. The person on the right remains still with no notable action."
- "There are 2 people in the scene: the person on the left (speaker 1), the person on the right (speaker 2). [0s–4s] The person on the left speaking while waving his hand. [2s–5s] The person on the right smiling with joy."

Action Categories. Each annotated event is associated with a free-form action description in our datasets. To enable systematic evaluation for action accuracy, we map these descriptions to one of 11 coarse action categories via keyword matching. Each category corresponds to a fixed yes/no question template used by the VLM evaluator, as listed in Table 7.

Table 7: Action categories and their VLM question templates. Each action accuracy query follows the format: "In this video, does {positional description} {template}? Answer Yes or No."

#	Category	Question template fragment
1	pointing	"point at someone or something"
2	object interaction	"interact with an object (e.g. pick up, put down, hold)"
3	head gesture	"make a head gesture (e.g. nod or shake head)"
4	mutual gesture	"physically interact with another person or clap"
5	body posture	"change body posture (e.g. cross arms, lean, stand up)"
6	speaking	"appear to be speaking (e.g. mouth moving)"
7	facial expression	"show a facial expression (e.g. smile or laugh)"
8	listening	"listen attentively"
9	looking	"look at someone or something"
10	hand gesture	"make a hand gesture (e.g. wave, raise hand)"
11	drinking/toasting	"drink or make a toast"

VLM Query Design. As introduced in Section 4.2, we formulate each metric as a binary VQA task. Each query is evaluated on a temporally cropped sub-clip $[\text{event.start} - 1\text{s},\ \text{event.end} + 1\text{s}]$. Speakers are identified via colored bounding boxes drawn on the video frames, and questions refer to "the person in the red/green box" rather than positional descriptions. Representative queries for each metric are listed below:

- **Action accuracy**: "In this video, does the person in the red box interact with any object? Answer Yes or No." (Expected: Yes)
- **Target accuracy**: "In this video, does the person in the red box point toward the person in the green box? Answer Yes or No." (Expected: Yes)
- **Stillness accuracy**: "In this video, does the person in the red box perform any notable gesture or directed action? Answer Yes or No." (Expected: No)

A.3 Reliability Analysis of Social Interaction Metrics

To verify that the improvements of SocialDirector reported in Section 5.3 are robust under our VLM-based evaluation pipeline rather than artifacts of a particular evaluator or random seed, we provide three complementary analyses: (i) per-VLM scores prior to majority voting, (ii) per-seed scores across three random seeds, and (iii) a human validation study of the VLM judges.

Per-VLM Consistency. Table 8 reports the social interaction metrics evaluated by each of the three VLMs (Qwen3-VL, InternVL3.5, MiniCPM-V) independently, before majority voting. Across all three evaluators, SocialDirector consistently improves over the Wan2.2 base model on all three social interaction metrics, indicating that the gains are not attributable to the bias of any single VLM.

Table 8: Per-VLM social interaction metrics (%) without majority voting. Green indicates improvement over the base model.

Method	Qwen3-VL			InternVL3.5			MiniCPM-V 4.5		
	Action ↑	Target ↑	Stillness ↑	Action ↑	Target ↑	Stillness ↑	Action ↑	Target ↑	Stillness ↑
GT Oracle	72.9	75.9	77.4	74.9	63.1	97.8	72.7	69.7	83.1
Wan2.2	69.8	73.9	73.5	72.8	61.7	95.0	70.2	67.6	74.6
Wan2.2 + SocialDirector	73.5 +3.7	77.3 +3.4	77.4 +3.9	75.7 +2.9	65.6 +3.9	95.5 +0.5	73.9 +3.7	69.7 +2.1	79.6 +5.0

Per-Seed Robustness. Table 9 reports the mean and standard deviation of social interaction metrics across three random seeds. On both base models, SocialDirector's gains on action and target accuracy clearly exceed the seed-level standard deviations, confirming that the improvements are statistically robust rather than seed-specific artifacts.

Table 9: Per-seed social interaction metrics (%, mean ± std over 3 seeds).

Method	Action Acc ↑	Target Acc ↑	Stillness Acc ↑
LTX-2.3	71.6 ± 1.3	69.4 ± 0.5	87.6 ± 0.9
LTX-2.3 + SocialDirector	74.3 ± 1.3	70.8 ± 0.5	88.3 ± 1.5
Wan2.2	72.2 ± 1.5	69.2 ± 1.6	85.1 ± 1.3
Wan2.2 + SocialDirector	76.1 ± 0.7	72.3 ± 0.8	88.2 ± 2.3

Human Validation. To verify that the open-source VLM judges are reliable for fine-grained social interaction assessment, we conduct a human validation study. Two annotators independently answer 72 sampled queries (24 per metric) over 10 real and 10 generated videos, given exactly the same trimmed clips, colored bounding boxes, and questions as the VLMs. Table 10 reports the agreement between humans, the VLM majority vote, and the benchmark annotations. The human study bounds what any judge can achieve on this task: the two annotators agree with each other only 80.6% of the time, and human answers match the benchmark annotation on real videos 86.1% of the time. Our VLM ensemble reaches 77.8% against the benchmark annotation and 70.8% against human answers, indicating that the automatic pipeline is close to human perception. Together with the per-VLM consistency above, this supports the conclusion that the gains of SocialDirector reflect genuine improvements in interaction fidelity rather than evaluator bias.

Table 10: Human validation of the VLM-based evaluation. N is the number of compared answer pairs.

Agreement	N	Accuracy / Agreement
Human answer vs. benchmark annotation (real videos)	72	86.1%
VLM vote vs. benchmark annotation (real videos)	36	77.8%
VLM vote vs. human answer (all queries)	144	70.8%
Human vs. human (all queries)	72	80.6%

A.4 Robustness to Noisy Inputs

SocialDirector relies on per-person bounding boxes and per-event time windows, which may be imprecise in practice. To examine how much this affects the method, we add two kinds of perturbations to the annotated inputs on Wan2.2: *noisy box*, where each bounding box is randomly enlarged or shrunk by up to 20%, and *noisy time*, where each event window is randomly rescaled by up to ±20% and its center randomly shifted by up to ±1s. These perturbations also introduce additional spatial and temporal overlap between events. As shown in Table 11, all accuracy drops remain below 3.5% under both perturbations, and target accuracy is unaffected by temporal noise, showing that moderate errors in boxes and timestamps, as well as the resulting overlap, are well tolerated. As discussed in Section 6, heavy overlap and occlusion remain a limitation, since a static box-based mask can no longer separate the actors.

Table 11: Robustness of SocialDirector (on Wan2.2) to perturbed bounding boxes and timestamps. Relative changes to the clean setting are shown in parentheses.

Input	Action Acc ↑	Target Acc ↑	Stillness Acc ↑
Clean boxes and timestamps	76.1	72.3	88.2
Noisy box (±20% scaling)	75.1 (−1.3%)	70.5 (−2.5%)	85.1 (−3.5%)
Noisy time (±20% scaling, ±1s shift)	75.3 (−1.1%)	72.3 (0.0%)	86.0 (−2.5%)

A.5 Comparison with Commercial Models

To contextualize the performance ceiling in this domain, we compare SocialDirector with a commercial state-of-the-art video generation model, Kling 3.0 Pro [30], on a 20-video subset of our benchmark. Kling receives the same first-frame image and text prompt as ours; since it exposes no interface for per-person bounding boxes or attention control, it serves as a reference for what a strong closed-source model achieves without explicit interaction control. As shown in Table 12, Kling leads on video quality metrics, reflecting its larger scale and training data. Nevertheless, SocialDirector surpasses Kling on action accuracy (75.4 vs. 72.9) and ties it on target accuracy (71.0),

while improving over the base Wan2.2 on every metric. This confirms that the interaction fidelity gained by SocialDirector is competitive even with commercial systems, and that explicit interaction control is complementary to raw generation quality.

Table 12: Comparison with the commercial model Kling 3.0 Pro on a 20-video subset of our benchmark.

Method	Social Interaction Metrics			Video Quality Metrics			
	Action Acc ↑	Target Acc ↑	Stillness Acc ↑	FVD ↓	LPIPS ↓	Vid Quality ↑	ID Pres ↑
GT Oracle	75.4	72.6	90.9	—	—	41.5	89.8
Wan2.2	72.0	67.7	81.8	4.61	18.3	46.3	85.8
Wan2.2 + SocialDirector	**75.4**	**71.0**	84.1	4.42	15.1	45.6	86.5
Kling 3.0 Pro	72.9	**71.0**	**93.2**	**4.18**	**13.0**	**51.3**	**87.2**

A.6 More Visualization Results

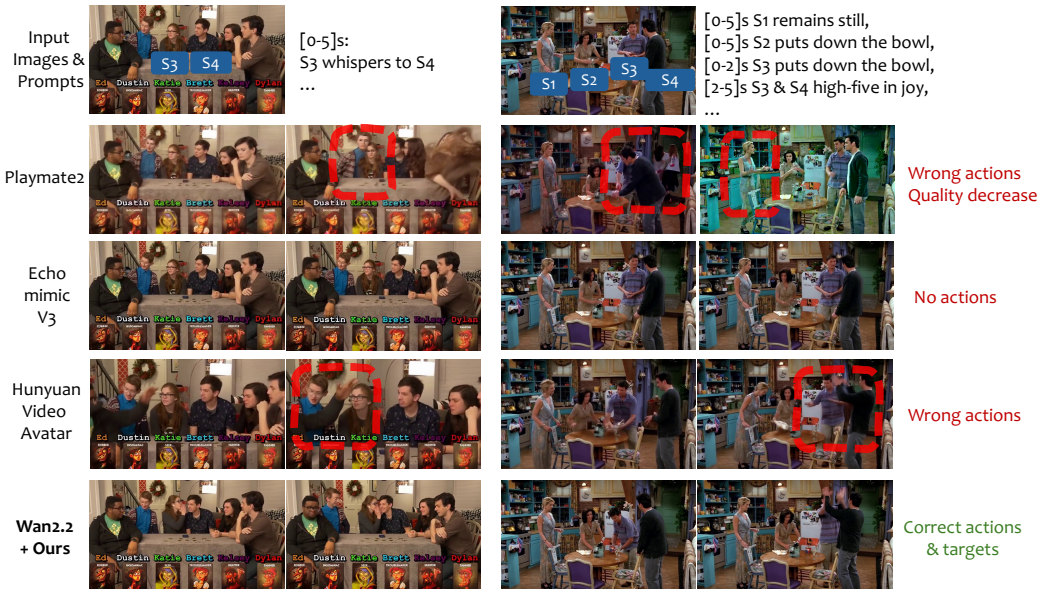


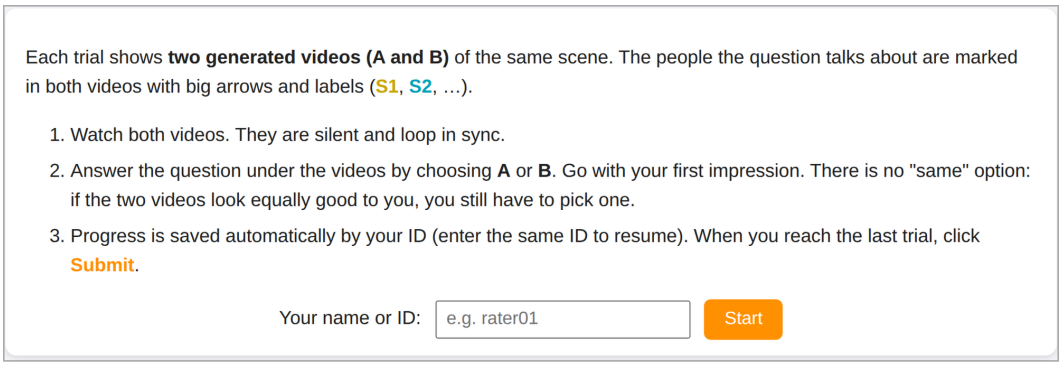
Figure 6: Additional qualitative comparisons of SocialDirector against baseline methods.

Figure 6 presents additional qualitative comparisons of SocialDirector against a broader set of baseline methods on diverse social interaction scenarios. Playmate2 [41] suffers from severe video quality degradation when applied to multi-person scenes with non-separated audio. EchoMimic v3 [45] produces nearly static videos with little interaction-relevant motion. HunyuanVideo-Avatar [10] likewise fails to handle multi-person scenes: in the left example it zooms in on S3 and S4, while in the right example it exhibits incorrect action–actor binding. Under the same inputs, SocialDirector correctly generates the intended interactions with accurate actor–action–target assignment and consistent identity across speakers.

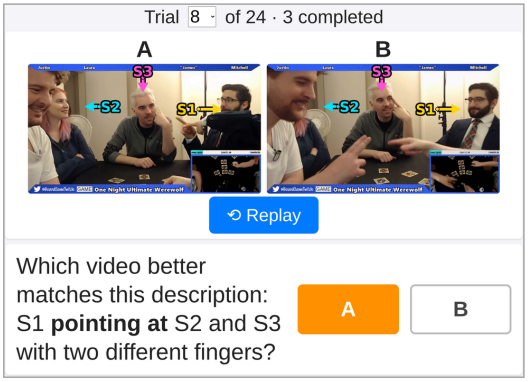
Video visualization. We provide a video compilation (`SocialDirector.mp4`) in the supplementary material that directly showcases the generation results of SocialDirector comparing with base models.

A.7 User Study Details

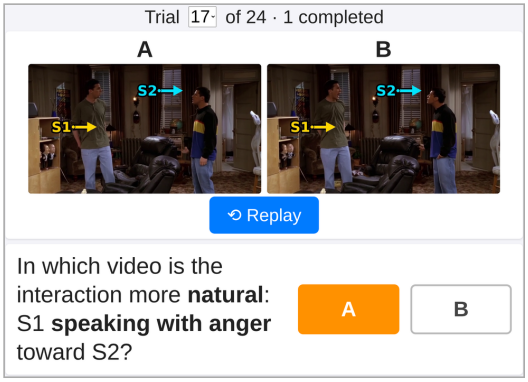
Figure 7 shows the user study interface. Each trial presents two silent videos generated from the same inputs (SocialDirector and a competing method, in random A/B order), with the persons mentioned in the question marked in both videos, and asks a single two-alternative forced-choice question on either description match or naturalness. Each participant completes 24 trials.



(a) Instruction page



(b) Description-match trial



(c) Naturalness trial

Figure 7: User study interface: (a) instruction page, (b) a description-match trial, and (c) a naturalness trial.

A.8 Broader Impacts

SocialDirector advances controllable multi-person video generation with potential positive impacts in film production, social robotics, virtual avatars, and accessibility applications such as creating engaging educational and communication content. As a human-centric video generation method, however, it also shares the broader risks of generative video technologies, including the possibility of being misused to synthesize misleading or ethically inappropriate content involving multiple individuals. Our research is conducted entirely on publicly available datasets and open-source pretrained models with no additional video data collection. To mitigate the risk of misuse, we preserve the built-in watermarks of the underlying base models (Wan2.2 and LTX-2.3) in all generated videos, so that outputs of our pipeline remain identifiable as synthetic content.

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 accurately reflect our main contributions: SocialDirector, a training-free cross-attention control framework for multi-person social interaction video generation, and the first fully automated VLM-based evaluation pipeline for social interaction video generation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in the Conclusion section, including (1) the lack of support for the audio modality, (2) the limited applicability to long video generation, (3) the reliance on structured inputs with timestamps and per-person bounding boxes, (4) the degraded control under heavy overlap, occlusion, and large actor movements, and (5) the possibility that VLM-based evaluation misjudges subtle social behaviors. Robustness to noisy inputs and human validation of the VLM judges are provided in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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 include any formal theoretical results, theorems, or proofs. It is an empirical work that proposes a training-free cross-attention control method and an evaluation pipeline, validated through experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

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 provides detailed descriptions of base models, method formulations, hyperparameters, baseline configurations, and evaluation protocols sufficient to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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: Our code and interaction annotations are released at `https://github.com/ut-vision/SocialDirector`. All datasets (MELD, MMSI, SocialGesture) and base models (Wan2.2, LTX-2.3) used in our experiments are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: All implementation and evaluation details are provided in Section 5 and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: [Yes]

Justification: Each experiment result is averaged over three runs with different random seeds to reduce the impact of randomness in video generation. We further report the per-seed mean and standard deviation of our main results in Table 9, where the seed-level standard deviations are clearly smaller than the gains of SocialDirector over its base models on action and target accuracy, confirming that the reported improvements are statistically robust rather than seed-specific artifacts. For the user study, we report bootstrap 95% confidence intervals over raters in Table 6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: Computation details are provided in Section 5.2 and Section 5.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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 research conforms with the NeurIPS Code of Ethics. We use only publicly available datasets and pretrained models. The user study only collects anonymous preference judgments on generated videos from voluntary participants.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We discuss both positive and negative societal impacts in Section A.8.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: [Yes]

    Justification: We preserve the built-in watermarks of the underlying base models (Wan2.2 and LTX-2.3) in all generated videos, ensuring that outputs remain identifiable as synthetic content. See Section A.8 for details.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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: All datasets (MELD, MMSI, SocialGesture) and pretrained models (Wan2.2, LTX-2.3) used in our paper are properly cited, and we comply with their respective licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We introduce new interaction-level annotations on top of existing datasets, which are released together with documentation at `https://github.com/ut-vision/SocialDirector`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: [Yes]

    Justification: Our user study (Section 5.5) involves 17 volunteer participants from The University of Tokyo and AlayaLab. Each participant was shown pairs of generated videos together with the interaction prompt and asked which video better matches the description and which looks more natural; the full instructions and screenshots of the interface are provided in Section A.7. No compensation was provided.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: [Yes]

    Justification: The user study only asks for subjective preferences between generated videos, poses no foreseeable risk to participants, and collects no personal information; IRB approval was therefore not required under our institutions' guidelines.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: [N/A]

    Justification: LLMs are not used as a component of the core method. They are only used for code review and writing assistance, which do not impact the methodology or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.