# *Multics*: GENERATIVE MULTI-PLAYER WORLD MODEL WITH HIERARCHICAL COMMUNICATION

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

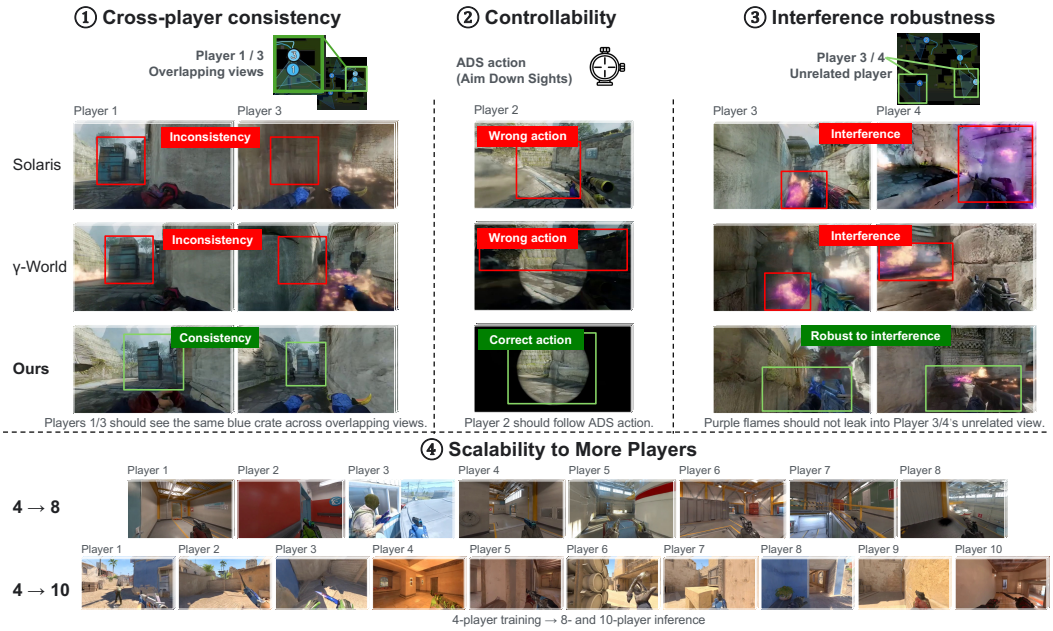



Figure 1: ***Multics* is a multi-player video generation model with hierarchical communication that structures information exchange across player, group, and global levels.** *Multics* can generate robust multi-player videos with cross-player consistency, player-specific controllability, and interference robustness, and can be generalized to more players without retraining.

## ABSTRACT

Multi-player interactive world models aim to simulate shared environments from multiple players' perspectives, but achieving robust generation while scaling to more players remains challenging. Existing approaches that rely on fully-connected communication (*e.g.*, Solaris) incur quadratic attention costs as the number of players grows, while centralized approaches (*e.g.*, $\gamma$-World) suffer from interference between unrelated events in a shared global representation. To address these challenges, we present *Multics*, a multi-player video generation framework with hierarchical communication that structures information exchange across player, group, and global levels. This hierarchy introduces group tokens into self-attention as intermediate group-level representations that aggregate locally relevant player information. These tokens connect to global tokens for cross-group communication, while the hierarchical topology limits direct mixing between unrelated local events. Group-token parameters are shared across group slots, allowing the same communication module to accommodate varying numbers of players without introducing new group-specific parameters. We further propose budget-constrained dynamic grouping based on recent camera trajectories, which adaptively groups relevant players while limiting unnecessary communication. Experiments demonstrate that *Multics* achieves state-of-the-art performance in 4- and 6-player settings, with improved cross-player consistency, player-specific action and camera controllability, and interference robustness. Moreover, *Multics* generalizes from 4-player training to 6-, 8-, and 10-player inference without retraining, while reducing self-attention computation compared with fully-connected communication.

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

# 1 INTRODUCTION

Multi-player interactive video world models aim to simulate shared environments from multiple players' perspectives. They have broad applications in game entertainment (Bruce et al., 2024) and interactive simulation (Ren et al., 2025; Fung et al., 2025). However, most video world models focus on single-player simulation (Wang et al., 2026; HunyuanWorld, 2025; Sun et al., 2025; Team et al., 2026) and generate observations conditioned on a single player's action, camera trajectory, and text prompt. Moving from single-player to multi-player simulation introduces two challenges: *robust multi-player generation* and *scalability to more players* (Figure 1). Robust multi-player generation requires generated observations to maintain cross-player consistency across overlapping views, faithfully follow player-specific action and camera controls, and remain robust to interference from unrelated concurrent events. Scalability to more players requires the model to generalize to larger player counts without retraining and scale efficiently as the number of players increases.

Existing multi-player video generation methods adopt different communication mechanisms for inter-player information exchange, but remain limited in jointly supporting robust generation and scalability. Fully-connected communication methods (Savva et al., 2026; Hu et al., 2026a) such as Solaris (Savva et al., 2026) enable unrestricted information exchange across players but incur quadratic computational cost as the player count grows. In contrast, centralized communication methods such as $\gamma$-World (Liu et al., 2026a) improve computational scalability through a shared global representation. However, routing heterogeneous interactions through the same communication pathway can mix information from unrelated events, increasing cross-event interference. This can further degrade cross-player consistency and player-specific controllability as concurrent interactions become more numerous and diverse. These limitations motivate structured local and global communication under a limited communication budget.

To address these challenges, we present *Multics*, a multi-player video generation framework with hierarchical communication that structures information exchange across the player, group, and global levels. Our key insight is that players vary in their relevance to one another, motivating a hierarchy that combines group-level aggregation of locally relevant information with global communication across groups. As shown in Figure 2, the hierarchy incorporates group tokens into self-attention as intermediate group-level representations that aggregate locally relevant player information and connect to global tokens for cross-group communication. This topology limits direct mixing between unrelated local events. Group-token parameters are shared across group slots, allowing the same communication module to accommodate varying numbers of players without introducing new group-specific parameters. To determine which players should share group-level representations, we propose budget-constrained dynamic grouping based on recent camera trajectories. This strategy adapts group assignments to player relevance while limiting unnecessary communication.

Experimental results show that our method achieves state-of-the-art performance in 4- and 6-player settings, with strong cross-player consistency, controllability, and interference robustness. Models trained with 4 players further generalize to 6-, 8-, and 10-player inference without retraining. Efficiency comparisons show substantially reduced self-attention computation relative to fully-connected communication. Ablation studies verify the effectiveness of our key designs. Our contributions can be summarized as follows:

- We introduce *Multics*, a multi-player video generation framework with hierarchical communication that structures inter-player information exchange across player, group, and global levels. Group tokens provide intermediate group-level representations for local aggregation while preserving global communication for cross-group exchange.
- We propose shared group-token parameters across group slots, enabling the same communication module to accommodate varying numbers of players without introducing new group-specific parameters. We further introduce budget-constrained dynamic grouping based on recent camera trajectories to organize relevant players while limiting unnecessary communication.
- Extensive experiments demonstrate state-of-the-art performance in multi-player generation, with strong cross-player consistency, player-specific action and camera controllability, and interference robustness. *Multics* further demonstrates generalization to larger player counts without retraining and maintains computational efficiency.

## 2 Related Work

**Video Generation and Video World Models.** Recent advances in video generation span Diffusion Transformers (DiTs) (Peebles & Xie, 2023; OpenAI, 2024), next-token and masked-token prediction (Villegas et al., 2022; Chang et al., 2023; Kondratyuk et al., 2023; Yu et al., 2023a;b; Chang et al., 2022), and hybrid methods (Chen et al., 2024; Deng et al., 2025), enabling high-quality video synthesis (DeepMind, 2024; Kling, 2025; Kong et al., 2024; Runway AI, 2025; Wang et al., 2025). Conditioning on camera trajectories (Bai et al., 2025; He et al., 2024; Wang et al., 2024b), actions (Yu et al., 2025b; Valevski et al., 2025), and multimodal references (MiniMax, 2024) further enables controllable generation. Building on the world-model paradigm of predicting future states from past states and actions (Ha & Schmidhuber, 2018), video world models serve as visual simulators by generating future observations (Bar et al., 2025; Liang et al., 2024; Yang et al., 2023b). These models have been explored in robotics (Li et al., 2026; Gao et al., 2026; Yang et al., 2023a; Ye et al., 2026), video games (Fan et al., 2022; Alonso et al., 2024; Wang et al., 2026; Valevski et al., 2025; Xiao et al., 2026), autonomous driving (Ren et al., 2025; Hu et al., 2023), and physical simulation (Yuan et al., 2026), demonstrating their potential as general-purpose visual simulators.

**Multi-Player World Models.** Recent multi-player world models use different communication mechanisms to maintain cross-player consistency. Dense joint-generation methods (Sun et al., 2026; Odyssey, 2026; Hu et al., 2026b), including Solaris (Savva et al., 2026) and MIRA (Hu et al., 2026a), couple player streams through shared or cross-agent attention, enabling fine-grained interactions with increasing computational cost as player counts grow. Centralized methods, including $\gamma$-World (Liu et al., 2026a) and MultiWorld (Wu et al., 2026), improve scalability by aggregating information through sparse hubs or compact global visual representations (Mo et al., 2026). Another line of work uses persistent or structured shared world states to decouple world-state evolution from per-player rendering (Po et al., 2026; Cai et al., 2026; Zhao et al., 2026a; Mo et al., 2026; Huang et al., 2026b). These shared-state methods primarily address population scalability through state-rendering decomposition. Our work explores an orthogonal design dimension: scalable communication within learned visual representations, using a hierarchical player-group-global topology to structure information exchange across representation levels.

**Communication Topologies.** Multi-agent communication includes dense exchange in CommNet (Sukhbaatar et al., 2016), selective or targeted communication in ATOC (Jiang & Lu, 2018) and TarMAC (Das et al., 2019), and hierarchical exchange across agent groups (Zhao et al., 2026b). While these methods primarily support multi-agent decision making, we study communication within high-dimensional video representations for multi-player world modeling.

## 3 Method

Given $N$ players in an environment, each player $i$ has an initial image $I_i \in \mathbb{R}^{c\times h\times w}$, a text prompt $P_i$, an action sequence $A_i \in \mathbb{R}^{(f-1)\times d_a}$, and a camera trajectory $C_i \in \mathbb{R}^{f\times 3\times 4}$, where $f, c, h, w, d_a$ denote the frame count, number of channels, height, width, and action dimension, respectively. Our goal is to jointly generate videos $\mathcal{V} = \{V_i\}_{i=1}^N$, $V_i \in \mathbb{R}^{f\times c\times h\times w}$, conditioned on $\{I_i, P_i, A_i, C_i\}_{i=1}^N$. The videos should maintain cross-player consistency when players observe overlapping views of the same scene or one another, and follow player-specific action and camera controls. Thus, we introduce hierarchical communication across the player, group, and global levels, as shown in Figure 2.

### 3.1 Multi-Player Video Generation with Conditioning

For each player $i$, the initial image $I_i$ serves as a clean anchor, while the action sequence $A_i$, camera trajectory $C_i$, and prompt $P_i$ condition the transformer blocks as detailed below. All players' visual tokens are concatenated along the sequence dimension for inter-player communication (Section 3.2).

**Action Conditioning.** We decompose $A_i = (D_i, U_i)$ into discrete controls $D_i \in \mathbb{R}^{(f-1)\times 23}$ and relative yaw/pitch changes $U_i \in \mathbb{R}^{(f-1)\times 2}$ (Savva et al., 2026). Following GameFactory (Yu et al., 2025b), we condition each latent frame on $rw_c$ action steps collected from a causal window spanning $w_c$ latent intervals. Here, $r$ is the video VAE's temporal compression ratio and $f_z = (f-1)/r + 1$ is the latent frame count. Discrete controls are projected to the transformer hidden dimension $d$, forming tokens in $\mathbb{R}^{f_z\times rw_c\times d}$ that are injected into the visual tokens via cross-attention. Mouse

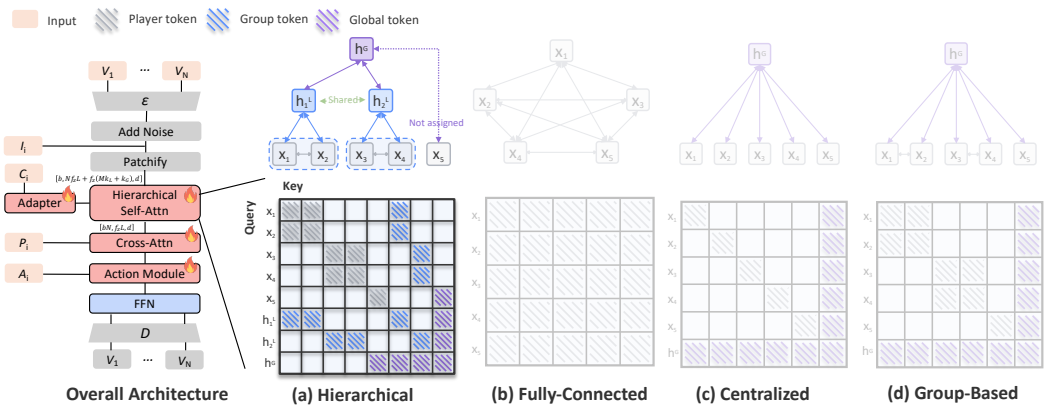



Figure 2: **Overview of *Multics*.** *Left*: The training pipeline of *Multics*. The video diffusion model is trained to jointly generate multi-player videos $\mathcal{V} = \{V_i\}_{i=1}^{N}$, conditioned on each player $i$'s initial image $I_i$, action sequence $A_i$, camera trajectory $C_i$, and text prompt $P_i$. *Right*: Comparison of different communication methods. **(a)** Hierarchical communication in our method; **(b)** Fully-connected communication in Savva et al. (2026); **(c)** Centralized communication in Liu et al. (2026a); **(d)** Group-based extension of centralized communication with direct intra-group player connections. The matrices illustrate attention connectivity at one latent timestep; block-causal temporal masking and spatial-token expansion are omitted for clarity.

controls are flattened into $\mathbb{R}^{f_z \times 2rw_c}$ features and fused with visual tokens through a lightweight temporal module.

**Camera and Prompt Conditioning.** Following ReCamMaster (Bai et al., 2025), we express all players' camera rotations and translations in a shared coordinate system anchored at the first player's initial frame, forming $\mathbf{C} \in \mathbb{R}^{N \times f \times 3 \times 4}$. Learnable camera adapters project $\mathbf{C}$ to the transformer hidden dimension and inject the resulting features into the corresponding player tokens' query and key features. For the group and global tokens introduced in Section 3.2, we use zero camera features as placeholders (Huang et al., 2026a). Text prompts are encoded and injected via cross-attention (Rombach et al., 2022).

### 3.2 Hierarchical Communication for Multi-Player

Effective multi-player generation requires structured information exchange across players. Let $\mathbf{x}_i \in \mathbb{R}^{f_z \times L \times d}$ denote the visual tokens of player $i$, where $L$ denotes the number of spatial tokens per frame. All communication topologies below are implemented using self-attention masks that define the allowed token-to-token connections. We review prior mechanisms (Savva et al., 2026; Liu et al., 2026a) and their natural extensions before introducing our hierarchical design.

**Fully-Connected Communication.** Solaris (Savva et al., 2026) concatenates the representations of all players along the sequence dimension and performs shared self-attention, allowing every player to communicate directly with all others:

$$\mathbf{x}_{\text{full}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]_{\text{seq}}, \qquad \mathbf{x}_i \leftrightarrow \mathbf{x}_j, \quad i, j = 1, \ldots, N. \tag{1}$$

We denote this topology as "fully-connected communication" and illustrate it in Figure 2(b).

**Centralized Communication.** Recent work (Liu et al., 2026a) introduces shared global hub tokens $\mathbf{h}^G \in \mathbb{R}^{f_z \times k_G \times d}$ as a common communication interface across players, where $k_G$ is the number of global tokens per latent frame. The communication sequence is

$$\mathbf{x}_{\text{cent}} = [\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{h}^G]_{\text{seq}}, \qquad \mathbf{x}_i \leftrightarrow \mathbf{h}^G, \quad i = 1, \ldots, N. \tag{2}$$

We denote this topology as "centralized communication" and illustrate it in Figure 2(c).

**Group-Based Communication.** As a straightforward extension, group-based communication introduces locality by restricting direct player-to-player communication to players within the same group. Let $g(i)$ denote the group assignment of player $i$. The resulting communication is

$$\mathbf{x}_i \leftrightarrow \mathbf{x}_j \quad \text{if} \quad g(i) = g(j), \qquad \mathbf{x}_i \leftrightarrow \mathbf{h}^G. \tag{3}$$

We denote this topology as "group-based communication" and illustrate it in Figure 2(d).

**Hierarchical Communication (Ours).** To improve cross-player consistency and player-specific controllability while mitigating interference among unrelated local events, we introduce group tokens as intermediate group-level representations into self-attention. Let $M$ denote the number of available group slots and $g \in \{1, \dots, M\}$ denote a group index. For each group slot $g$, we instantiate $k_L$ group tokens $\mathbf{h}_g^L \in \mathbb{R}^{f_z \times k_L \times d}$ between player- and global-level communication. Let $\mathbf{h}^L$ denote the collection of group tokens from all $M$ group slots. For player $i$ assigned to group $g(i)$, the hierarchical communication is formulated as

$$\mathbf{x}_{\text{hier}} = [\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{h}^L, \mathbf{h}^G]_{\text{seq}}, \qquad \mathbf{x}_i \leftrightarrow \mathbf{h}_{g(i)}^L \leftrightarrow \mathbf{h}^G. \tag{4}$$

Players within the same group can directly attend to one another. Players that are not assigned to a group communicate directly through the global tokens. We denote this topology as "hierarchical communication" and illustrate it in Figure 2(a).

To avoid binding group representations to fixed group identities, we learn a set of shared group-token parameters $\mathbf{E}^L \in \mathbb{R}^{k_L \times d}$. At runtime, $\mathbf{E}^L$ is replicated across the $M$ group slots and broadcast over latent frames to instantiate $\mathbf{h}^L \in \mathbb{R}^{f_z \times M \times k_L \times d}$. Although all group slots share the same learnable parameters, their hidden representations evolve through hierarchical self-attention on the players assigned to each group, yielding group-specific representations. This design of shared group-token parameters makes the group slots exchangeable and enables the same communication module to accommodate varying numbers of players, without introducing new group-specific parameters or retraining the model. Appendix C.2 reports an ablation of shared group-token parameters.

**Comparison and Discussion.** Our experimental results demonstrate the substantial advantages of hierarchical communication over fully-connected, centralized, and group-based alternatives, as illustrated in Figure 3 and Figure 5. We attribute these gains to the multi-granularity representations enabled by hierarchical communication. By introducing intermediate group-level representations between player- and global-level communication, our method allows locally related information to be aggregated before broader global exchange, reducing the direct mixing of heterogeneous information, such as actions, visual details, and unrelated local events, within a shared communication space. This facilitates more effective multi-player representation learning and improves cross-player consistency, player-specific action and camera controllability, and interference robustness.

### 3.3 Dynamic Grouping

Hierarchical communication requires determining which players should share group-level representations. Treating all players as equally relevant for group communication introduces unnecessary information exchange and computation and can increase cross-event interference. We therefore introduce dynamic grouping, which estimates pairwise player relevance from recent camera trajectories (Yu et al., 2025a) and organizes relevant players into groups under a communication budget.

**Camera-Based Player Relevance.** Building on the camera-based field-of-view (FoV) overlap in Context-as-Memory (Yu et al., 2025a), we estimate player relevance from recent camera trajectories. At each grouping update $t$, we compute a symmetric matrix $\mathbf{R}^{(t)} \in [0,1]^{N \times N}$ from observed camera histories: $R_{ij}^{(t)} = \text{Rel}(C_i^{\leq t}, C_j^{\leq t})$, where Rel measures spatial proximity and viewing-direction alignment using only past and current camera states to preserve causality. Higher scores indicate that the corresponding players are more likely to observe spatially related content and benefit from local exchange. Relevance determines only the communication topology, while feature exchange is learned through hierarchical self-attention. Details of computation are provided in Appendix B.

**Budget-Constrained Dynamic Grouping.** Given $N$ players, we determine the number of group slots $M$ from the player count and derive the maximum group capacity $K$ under a communication budget $\mathcal{C}_{\max}$ that matches the communication cost of the centralized baseline. At time $t$, the active groups $\mathcal{G}^{(t)}$ must satisfy $|\mathcal{G}^{(t)}| \leq M$, $|\mathcal{S}| \leq K$ for all $\mathcal{S} \in \mathcal{G}^{(t)}$, and $\mathcal{C}(\mathcal{G}^{(t)}) \leq \mathcal{C}_{\max}$, where $\mathcal{S}$ is a set of players assigned to one group and $\mathcal{C}$ measures the communication cost of the induced attention topology. Using $\mathbf{R}^{(t)}$, we select a feasible grouping that maximizes player coverage while preserving high-relevance player relations. Remaining group slots and budget are used to accommodate unassigned players; otherwise, they communicate only through global tokens. Details of the determination of $M$, budget calculation, and derivation of $K$ are provided in Appendix B.

# 4 EXPERIMENTS

## 4.1 EXPERIMENT SETTING

**Implementation Details.** We train *Multics* based on Wan2.2 TI2V 5B (Wang et al., 2025) using OpenCS2 (Blanchon, 2026), an open multiplayer dataset rendered from ten-player Counter-Strike 2 matches with camera trajectories and mouse/keyboard action annotations. We train separate models for $N \in \{4, 6\}$, randomly selecting $N$ of the ten players for each sample. Each sample contains $N$ videos of 165 frames at $352 \times 640$ resolution. We adopt diffusion forcing training (Chen et al., 2024) and update the dynamic grouping topology every 48 video frames using causal camera trajectories. Additional details are provided in Appendix A.

**Baselines.** Given the lack of existing open-source methods for evaluating more than 4 players, we adopt the following baselines. (1) *Fully-Connected Communication*: Solaris (Savva et al., 2026) pioneered multi-player video generation by concatenating all players' tokens along the sequence dimension for shared self-attention. (2) *Centralized Communication*: $\gamma$-World (Liu et al., 2026a) introduces a set of shared hub tokens for information aggregation across players. For a fair comparison, we reimplement all baselines using the same base model and dataset, with identical training configurations and iterations.

**Evaluation Metrics.** We employ 10 metrics across four key aspects: consistency, action accuracy, camera accuracy, and visual quality. (1) *Consistency*: following prior work (Wang et al., 2024a; Liu et al., 2026b), we evaluate cross-view consistency with GeoKPM, which uses geometry-aware keypoint matching under epipolar-geometry filtering. Following Liu et al. (2026a), we also report the pixel-aligned metrics PSNR and SSIM (Wang et al., 2004) and the perceptual metric LPIPS (Zhang et al., 2018). (2) *Action accuracy*: following prior work (Baker et al., 2022; Wu et al., 2026), we use a trained Inverse Dynamics Model (IDM) to infer actions from generated videos. (3) *Camera accuracy*: we use RotErr, TransErr, and CamMC, following prior work (Zheng et al., 2024; Huang et al., 2026a). (4) *Visual quality*: following prior work (Savva et al., 2026; Liu et al., 2026a), we evaluate generation quality with FID (Heusel et al., 2017) and FVD (Unterthiner et al., 2018). For evaluation, following Yu et al. (2025a), we randomly select 200 held-out OpenCS2 samples, with 100 for each separately trained model. To assess cross-dataset applicability, we also train on EgoTeam's training set (Peng et al., 2026), a dataset of multi-robot egocentric videos rendered in reconstructed real-world indoor environments, and evaluate on 100 test samples.

## 4.2 COMPARISON WITH PREVIOUS METHODS

**Quantitative Results.** We compare *Multics* with Solaris and $\gamma$-World in Table 1 (top). Our method outperforms the baselines on most evaluation metrics in both 4- and 6-player settings, demonstrating strong cross-player consistency, action/camera controllability, and visual quality. Despite unrestricted cross-player information exchange, Solaris underperforms *Multics* on most metrics while incurring substantial computational overhead, suggesting that dense communication does not necessarily translate into better multi-player generation. $\gamma$-World struggles with consistency and camera accuracy due to its reliance on the same shared global pathway, which aggregates heterogeneous player information, leading to degraded cross-player consistency and player-specific controllability under diverse concurrent interactions. We attribute *Multics*'s gains to hierarchical communication, which introduces intermediate group-level representations to effectively structure multi-player information exchange. As pixel-aligned metrics PSNR and SSIM compare pixel values or local image statistics at corresponding spatial locations, they are sensitive to small spatial deviations, whereas the consistent gains in GeoKPM and LPIPS indicate stronger geometric and perceptual cross-player consistency of our method.

**Qualitative Results.** As shown in Figure 3, *Multics* demonstrates stronger cross-player consistency, controllability, and interference robustness. Solaris (Savva et al., 2026) fails to preserve consistency across overlapping views of the same scene or one another, while $\gamma$-World (Liu et al., 2026a) is more susceptible to information mixing through its shared global pathway, causing inconsistent visual details and player-specific actions being affected by unrelated observations. Since all methods share the same video-generation backbone, the results highlight the effectiveness of hierarchical communication for robust multi-player generation.

Table 1: **Quantitative comparison with previous methods**. We compare *Multics* with Solaris (Savva et al., 2026) and $\gamma$-World (Liu et al., 2026a) on consistency, action accuracy, camera accuracy, and visual quality. *Top*: Results with 4 and 6 players. *Middle*: Generalization to more players without additional training, where $4 \rightarrow N$ denotes inference with $N$ players using models trained with 4 players. *Bottom*: Results on EgoTeam. Bold numbers indicate the best result within each setting.

| Method | Consistency | | | | Action Accuracy | Camera Accuracy | | | Visual Quality | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GeoKPM↑ | PSNR↑ | SSIM↑ | LPIPS↓ | IDM↑ | RotErr↓ | TransErr↓ | CamMC↓ | FID↓ | FVD↓ |
| **Top: Quantitative Comparison** | | | | | | | | | | |
| *4 Players* | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 48.51% | 17.62 | 0.4952 | 0.4324 | 93.87 | 2.572 | 13.50 | 15.00 | 28.41 | 171.9 |
| $\gamma$-World (Liu et al., 2026a) | 49.42% | 17.76 | **0.5009** | 0.4293 | 93.98 | 2.728 | 13.66 | 15.26 | 29.36 | 200.5 |
| **Ours** | **54.17%** | **18.10** | 0.5005 | **0.4184** | **94.17** | **2.434** | **11.67** | **13.21** | **26.39** | **162.3** |
| *6 Players* | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 54.65% | 17.11 | **0.4728** | 0.4467 | 94.32 | 3.991 | 11.38 | 14.19 | 18.83 | 115.1 |
| $\gamma$-World (Liu et al., 2026a) | 52.80% | 17.21 | 0.4721 | 0.4389 | **94.34** | 3.761 | 11.59 | 14.29 | 20.16 | 128.4 |
| **Ours** | **56.84%** | **17.33** | 0.4687 | **0.4311** | 94.33 | **3.669** | **10.49** | **13.21** | **16.34** | **101.2** |
| **Middle: Generalization to More Players** | | | | | | | | | | |
| *4 → 6 Players* | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 49.18% | 16.58 | 0.4674 | 0.4689 | 94.17 | 3.979 | 12.22 | 14.86 | 23.16 | 148.2 |
| $\gamma$-World (Liu et al., 2026a) | 48.50% | 16.74 | **0.4695** | 0.4618 | 94.20 | 4.017 | 12.87 | 15.58 | 24.90 | 184.3 |
| **Ours** | **54.33%** | **16.97** | 0.4659 | **0.4517** | **94.31** | **3.760** | **11.15** | **13.74** | **21.80** | **133.8** |
| *4 → 8 Players* | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 50.37% | 16.38 | 0.4703 | 0.4778 | 94.16 | 3.932 | 12.40 | 15.08 | 21.47 | 133.8 |
| $\gamma$-World (Liu et al., 2026a) | 49.17% | 16.65 | **0.4712** | 0.4664 | 94.19 | 3.794 | 11.75 | 14.53 | 23.45 | 172.3 |
| **Ours** | **51.05%** | **16.82** | 0.4676 | **0.4600** | **94.26** | **3.553** | **11.52** | **13.92** | **20.07** | **117.0** |
| *4 → 10 Players* | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 47.82% | 16.06 | **0.4670** | 0.4921 | 93.95 | 7.320 | 12.91 | 18.68 | 21.28 | 139.0 |
| $\gamma$-World (Liu et al., 2026a) | 49.12% | 16.38 | 0.4667 | 0.4741 | 94.11 | 7.138 | 12.26 | 17.95 | 22.37 | 154.9 |
| **Ours** | **51.54%** | **16.52** | 0.4623 | **0.4704** | **94.17** | **5.941** | **11.42** | **16.14** | **19.23** | **121.9** |
| **Bottom: Evaluation on EgoTeam** | | | | | | | | | | |
| Solaris (Savva et al., 2026) | 57.65% | **16.44** | 0.5673 | 0.4381 | 96.76 | 0.3162 | 1.247 | 1.422 | 38.91 | 228.9 |
| $\gamma$-World (Liu et al., 2026a) | 52.31% | 16.18 | 0.5659 | 0.4421 | 97.04 | 0.3057 | 1.239 | 1.398 | 36.09 | 196.4 |
| **Ours** | **58.62%** | 16.33 | **0.5710** | **0.4373** | **97.07** | **0.2894** | **1.232** | **1.381** | **34.18** | **195.1** |

Table 2: **Ablation on grouping strategies.**

| Method | Consistency | | | | Action Accuracy | Camera Accuracy | | | Visual Quality | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GeoKPM↑ | PSNR↑ | SSIM↑ | LPIPS↓ | IDM↑ | RotErr↓ | TransErr↓ | CamMC↓ | FID↓ | FVD↓ |
| Group-based | 48.26% | 17.50 | 0.4878 | 0.4317 | 94.02 | 2.780 | 13.18 | 14.91 | 28.76 | **158.5** |
| Fixed grouping | 50.71% | **18.16** | **0.5052** | 0.4197 | 94.05 | 2.496 | 12.09 | 13.71 | 27.78 | 168.0 |
| Random grouping | 46.76% | 17.78 | 0.5036 | 0.4329 | 93.50 | 4.796 | 14.61 | 17.90 | 29.83 | 208.2 |
| **Ours** | **54.17%** | 18.10 | 0.5005 | **0.4184** | **94.17** | **2.434** | **11.67** | **13.21** | **26.39** | 162.3 |

### 4.3 Scalability and Generalization Ability of *Multics*

**Computational Efficiency.** Following Liu et al. (2026a), we compare communication methods using DiT latency, self-attention latency, and self-attention FLOPs. As shown in Figure 6, fully-connected communication scales quadratically with player count, while *Multics* substantially reduces computation (9.6 vs. 69.1 TFLOPs with 10 players) and remains comparable to centralized and group-based methods. These results demonstrate the computational efficiency of our method for scalable multi-player world models.

**Generalization to More Player Counts.** We apply models trained with 4 players to 6-, 8-, and 10-player inference without retraining or introducing additional model parameters. The results in Table 1 (middle) highlight *Multics*'s ability to generalize to larger player counts while achieving stronger consistency, controllability, and visual quality than Solaris and $\gamma$-World.

**Cross-Domain Applicability.** Beyond game environments, we also evaluate *Multics* on EgoTeam (Peng et al., 2026). As shown in Table 1 (bottom), our method maintains robust multi-player generation performance, demonstrating its cross-domain applicability.

### 4.4 Ablation Study

**Effectiveness of Hierarchical Communication.** We train four models with different communication schemes on the 4-player OpenCS2 data: fully-connected, centralized, group-based, and hierar-

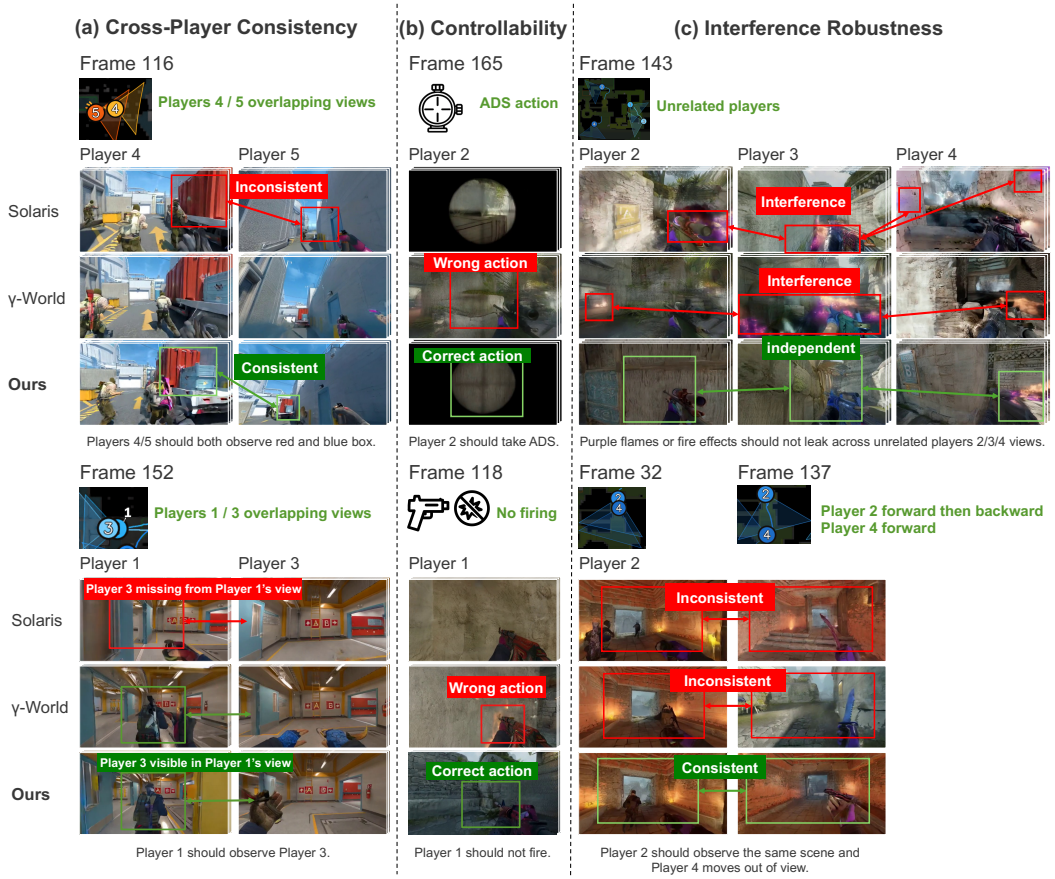



Figure 3: **Qualitative comparison of *Multics* and fully-connected and centralized baselines in 4-/6-player settings.** We compare *Multics* with Solaris (Savva et al., 2026) and $\gamma$-World (Liu et al., 2026a) on **(a)** cross-player consistency: players observing overlapping views of the same scene and one another; **(b)** player-specific controllability: adherence to aim-down-sights (ADS) and no-firing controls; and **(c)** interference robustness: independence of unrelated concurrent events and preservation of a player's scene upon revisiting despite other players' movements. Top-down maps are provided for visualization only, and are not used as conditioning inputs. Other players are omitted for clarity, and can be found in supplementary videos.




Figure 4: **Qualitative comparison of *Multics* and fully-connected and centralized baselines in generalizing to 6-/8-/10-player scenarios.**

chical communication (ours). Results in Table 1 (top) and Table 2 show that *Multics* consistently improves cross-player consistency and action/camera controllability. In particular, centralized communication is more susceptible to heterogeneous information mixing through its shared global representation, while group-based communication (Figure 5) restricts local communication but lacks an

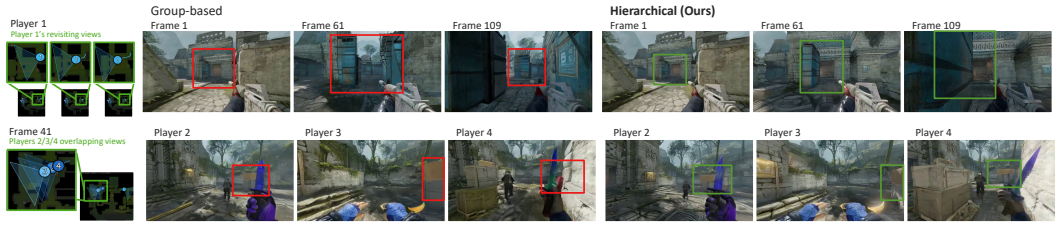



Figure 5: **Qualitative ablation study on communication mechanisms.** The group-based method exhibits inconsistencies when revisiting previously observed scenes (top) and across player views (bottom).

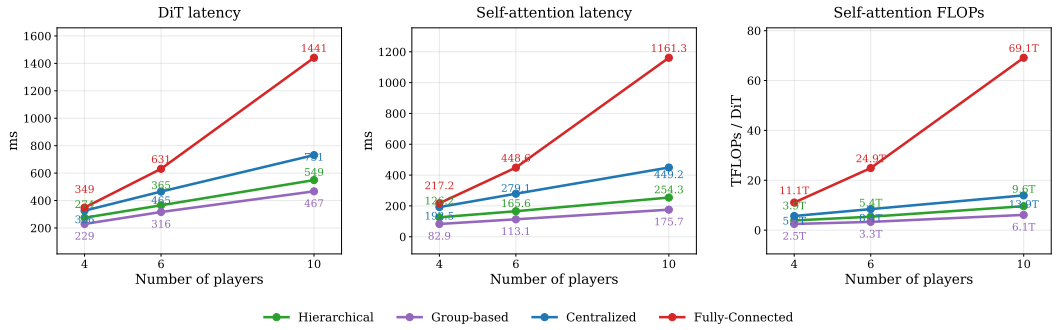



Figure 6: **Comparisons of computational efficiency.** *Multics* reduces computation relative to fully-connected communication and achieves efficiency comparable to centralized and group-based methods.

intermediate group-level representation. These results support the effective use of representations at multiple granularities to structure multi-player information exchange.

**Impact of Dynamic Grouping.** We train three sets of models with different grouping strategies on 4-player OpenCS2 data: (1) fixed grouping; (2) random grouping; (3) dynamic grouping (ours). As shown in Table 2, fixed grouping yields lower GeoKPM scores and reduced action and camera accuracy compared to dynamic grouping. Random grouping also leads to substantial decreases in GeoKPM and camera accuracy. These results demonstrate that dynamic grouping improves consistency and controllability by enabling relevant information exchange among multiple players.

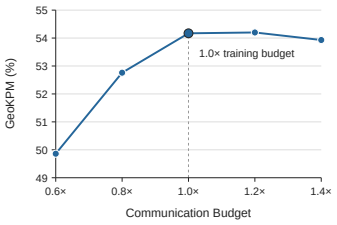



Figure 7: Effect of communication budget on GeoKPM.

**Sensitivity to Communication Budget.** We evaluate how changes in the inference communication budget affect cross-player geometric consistency. We train the model with the default budget of $1.0\times$ and vary the inference budget from $0.6\times$ to $1.4\times$ without retraining. As shown in Figure 7, reducing the budget below $1.0\times$ leads to a decrease in GeoKPM, whereas increasing it beyond $1.0\times$ leaves GeoKPM relatively stable.

## 5 Conclusion

We present *Multics*, a multi-player video generation framework with hierarchical communication that structures information exchange across player, group, and global levels. We introduce group tokens as group-level representations into self-attention, aggregating locally relevant player information while retaining global communication for cross-group exchange. We propose shared group-token parameters across group slots, allowing the same communication module to accommodate varying player counts without introducing new group-specific parameters. We further introduce budget-constrained dynamic grouping based on recent camera trajectories to determine which players share group-level representations. Extensive experiments demonstrate *Multics*'s state-of-the-art performance in generating robust multi-player videos with cross-player consistency, controllability, and interference robustness, while scaling to more players.

### AI USE STATEMENT

We used generative AI tools to assist with language editing, organization of the manuscript, refinement of figure captions, and identification of potentially relevant references. The authors reviewed and revised the AI-assisted text, verified cited references and their relevance to the associated claims, and checked technical descriptions against the implemented methods and experimental results. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

### ETHICS STATEMENT

This work studies multi-player video generation using existing game-rendered and simulated robot datasets. Generated videos may contain inaccuracies or be misused to create misleading content. We encourage clear disclosure of synthetic content and appropriate validation before use in real-world applications.

### REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide the detailed descripitions of methodology in Section 3 and the Appendix.

## REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801. IEEE, 2025.

Julien Blanchon. Opencs2 dataset. https://github.com/julien-blanchon/opencs2-dataset, 2026.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Ziqi Cai, Siqi Yang, Yimu Wang, Zixian Gao, Yunheng Liu, Shuchen Weng, Erwin Wu, Kaipeng Zhang, and Boxin Shi. Mass: Multiplayer world models with authoritative shared state. *arXiv preprint arXiv:2608.06257*, 2026.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 1538–1546. PMLR, 2019.

DeepMind. Video generation model with expanded creative controls. https://deepmind.google/models/veo/, 2024.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, volume 2025, pp. 44730–44745, 2025.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35: 18343–18362, 2022.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Alyx Liao, Amélie Royer, Manu Orsini, Adam Jelley, Eloi Alonso, Florian Laurent, Fredrik Norén, James Swingos, Jan Hünermann, Kent Rollins, Lucas Hosseini, Matthieu Le Cauchois, Maxim Peter, Pim de Witte, Tim Brown, Vincent Micheli, Moritz Böhle, Gabriel de Marmiesse, Viktoriia Sharmanska, Lucia Specia, Michael Black, and Patrick Pérez. MIRA: Multiplayer interactive world models with representation autoencoders, 2026a. Technical Report.

Li Hu, Guangyuan Wang, Peng Zhang, and Bang Zhang. Wantofight: Real-time generative game engine for multi-player combat interaction. *arXiv preprint arXiv:2607.12592*, 2026b.

Kaiyi Huang, Yukun Huang, Yu Li, Jianhong Bai, Xintao Wang, Zinan Lin, Xuefei Ning, Jiwen Yu, Yu Wang, and Xihui Liu. Cinescene: Implicit 3d as effective scene representation for cinematic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25381–25392, 2026a.

Zheng-Hui Huang, Guixu Lin, Jiacheng Lin, Yi-Chuan Huang, Ruihan Yu, Muyao Niu, Siqi Yang, Yu-Lun Liu, Yung-Yu Chuang, Kaipeng Zhang, et al. Programmable world model. *arXiv preprint arXiv:2609.10540*, 2026b.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

Jiechuan Jiang and Zongqing Lu. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7265–7275, 2018.

Kling. https://kling.kuaishou.com/, 2025. Organization: Kuaishou.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Junkun Yuan, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yanxin Long, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024. URL https://arxiv.org/abs/2412.03603.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

Fangfu Liu, Kai He, Tianchang Shen, Tianshi Cao, Sanja Fidler, Yueqi Duan, Jun Gao, Igor Gilitschenski, Zian Wang, and Xuanchi Ren. Gamma-world: Generative multi-agent world modeling beyond two players. *arXiv preprint arXiv:2605.28816*, 2026a.

Lulin Liu, Nuo Chen, Yan Wang, Bangya Liu, Wenyan Cong, Hezhen Hu, Boris Ivanovic, Hao Wang, Ziyao Zeng, Xinyu Gong, et al. Openlongtail: Generative scaling of long-tail driving data. *arXiv preprint arXiv:2607.09655*, 2026b.

MiniMax. MiniMax H3: An Open Model Breaking the Boundaries Between Tasks and Modalities. https://openai.com/index/sora/, 2024.

Sicheng Mo, Yuheng Li, Ziyang Leng, Krishna Kumar Singh, and Bolei Zhou. Streaming multi-agent autoregressive diffusion model with world state registers. *arXiv preprint arXiv:2607.21594*, 2026.

Odyssey. Agora-1: The multi-agent world model. https://odyssey.ml/introducing-agora-1, May 2026. Research preview.

OpenAI. Creating video from text. https://openai.com/index/sora/, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Kunyu Peng, Zhikun Zhou, Kailun Yang, Di Wen, Ruiping Liu, Yufan Chen, Junwei Zheng, Hao Shi, Yi Zhou, M Saquib Sarfraz, et al. Seeing together: Multi-robot cooperative egocentric spatial reasoning with multimodal large language models. *arXiv preprint arXiv:2605.18431*, 2026.

Ryan Po, David Junhao Zhang, Amir Hertz, Gordon Wetzstein, Neal Wadhwa, and Nataniel Ruiz. Multigen: Level-design for editable multiplayer worlds in diffusion game engines. *arXiv preprint arXiv:2603.06679*, 2026.

Xuanchi Ren, Yifan Lu, Tianshi Cao, Ruiyuan Gao, Shengyu Huang, Amirmojtaba Sabour, Tianchang Shen, Tobias Pfaff, Jay Zhangjie Wu, Runjian Chen, et al. Cosmos-drive-dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Runway AI. Gen-4. https://runwayml.com/research/introducing-runway-gen-4, 2025.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*, volume 29, pp. 2244–2252, 2016.

Huiqiang Sun, Zhan Peng, Size Wu, Kun Wang, Kang Liao, Dianyi Wang, Xingyu Zeng, Sheng Jin, Yangguang Li, Zhiguo Cao, Ziwei Liu, and Wei Li. Prisma-world: Camera-controllable multi-agent video world model. *arXiv preprint arXiv:2606.09507*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world model. *arXiv preprint*, 2025.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

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.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14749–14759. IEEE, 2024a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. Multiworld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023a.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1389–1399. IEEE, 2023b.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599. IEEE, 2025b.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023a.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023b.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

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*, pp. 586–595, 2018.

Renjie Zhao, Yuxiang Wu, Mingyu Zhang, Jiaxin Li, Sisi Li, Yimin Sheng, Tianxi Tan, Zhenkai Zhang, Jianyi Zhu, and Yong-Lu Li. Population-scalable multi-agent world modeling. *arXiv preprint arXiv:2608.08600*, 2026a.

Runze Zhao, Dongruo Zhou, Sumit Kumar Jha, Nathaniel D. Bastian, and Ankit Shah. Hicomm: Hierarchical communication for multi-agent reinforcement learning. *arXiv preprint arXiv:2606.29126*, 2026b.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

# A Implementation Details

We adopt diffusion forcing training (Chen et al., 2024), with a causal block of 3 latent frames. We train the model for 10K steps with a local attention window of 7 latent frames, followed by 25K steps with a window of 21 latent frames, using a learning rate of $3 \times 10^{-5}$. We use AdamW for the two-stage training schedule described above. The per-GPU batch size is one multi-player sample, and we train on 16 GPUs, resulting in an effective batch size of 16 multi-player samples. Training is performed in bfloat16 with gradient checkpointing and a gradient clipping norm of 1.0. We freeze the VAE, text encoder, feed-forward layers, and the remaining pretrained DiT parameters, while optimizing the self-attention and prompt cross-attention layers, action control modules, camera Q/K adapters, and the learnable group and global token embeddings. We apply a dropout rate of 0.1 to both the camera and text conditions. During inference, we generate 165-frame videos at a resolution of $352 \times 640$ using 50 denoising steps and a classifier-free guidance scale of 5.0. For hierarchical communication, we use 5 group tokens for each group and 6 global tokens. We apply simplex rotary encoding for players (Liu et al., 2026a) to our method and all baselines to make all players permutation-equivariant.

# B Details of Dynamic Grouping

**Camera-Based Player Relevance.** At each grouping step $t$, we estimate pairwise player relevance using only camera states observed up to the current time, ensuring strict causality. Given the recent camera trajectories of $N$ players, we compute a symmetric relevance matrix

$$\mathbf{R}^{(t)} \in [0,1]^{N \times N}, \tag{5}$$

where $R_{ij}^{(t)}$ measures the geometric relevance between players $i$ and $j$.

Following the camera-based relevance principle in prior work (Yu et al., 2025a), we combine two complementary geometric cues. First, we estimate *scene relevance* from camera proximity and viewing direction. Let $\mathbf{p}_i(\tau)$ and $\mathbf{v}_i(\tau)$ denote the camera position and unit viewing direction of player $i$ at frame $\tau$. We compute a proximity score from the camera distance

$$r_{ij}^{\text{prox}}(\tau) = \exp\left(-\frac{\|\mathbf{p}_i(\tau) - \mathbf{p}_j(\tau)\|_2}{\sigma_p}\right), \tag{6}$$

and estimate viewing similarity by projecting each camera direction to a fixed-distance endpoint $\mathbf{e}_i(\tau) = \mathbf{p}_i(\tau) + \ell \mathbf{v}_i(\tau)$:

$$r_{ij}^{\text{view}}(\tau) = \exp\left(-\frac{\|\mathbf{e}_i(\tau) - \mathbf{e}_j(\tau)\|_2}{\sigma_v}\right). \tag{7}$$

Their temporally aggregated combination gives the scene relevance $R_{ij}^{(t),\text{scene}}$.

Second, we estimate *mutual-view relevance* according to whether the players fall inside each other's camera fields of view. Let $Q_{i\to j}(\tau) \in [0,1]$ denote the pose-based visibility of player $j$ in player $i$'s view. We combine the two directed visibility scores as

$$r_{ij}^{\text{mutual}}(\tau) = \sqrt{Q_{i\to j}(\tau) Q_{j\to i}(\tau)}, \tag{8}$$

and aggregate it over the recent causal history to obtain $R_{ij}^{(t),\text{mutual}}$.

We use recency-weighted temporal aggregation and smoothing to make the relevance robust to short-term camera fluctuations. The final pairwise relevance is

$$R_{ij}^{(t)} = \max\left(R_{ij}^{(t),\text{scene}}, R_{ij}^{(t),\text{mutual}}\right). \tag{9}$$

Higher values indicate that the corresponding players are more likely to share locally relevant visual information. Importantly, the relevance matrix determines only the communication topology; the actual information exchanged between connected representations is learned through self-attention.

**Communication Budget.** Given $N$ players, we determine the number of available group slots $M$ using a slowly growing function of the player count. In our implementation, we use

$$M = \lceil \log_2 N \rceil. \tag{10}$$

We constrain hierarchical communication by the communication cost of the centralized baseline. Ignoring the same-player temporal-history term shared by both topologies, the centralized communication cost is

$$\mathcal{C}_{\text{central}} = A_w \left(2NLk_C + k_C^2\right), \tag{11}$$

where $L$ denotes the number of visual tokens per player per latent frame, $k_C$ the number of centralized hub tokens, and $A_w$ the number of allowed temporal frame pairs.

To derive a preliminary grouping capacity, we use a reference topology with all local hubs active, each player attached to exactly one local hub, and no direct player-to-player communication. Its reference cost is

$$\mathcal{C}_{\text{hier-base}} = A_w \left(2NLk_L + Mk_L^2 + 2Mk_Lk_G + k_G^2\right), \tag{12}$$

where $k_L$ and $k_G$ denote the numbers of group and global communication tokens per latent frame, respectively. The remaining communication budget is therefore

$$\Delta\mathcal{C} = \mathcal{C}_{\text{central}} - \mathcal{C}_{\text{hier-base}}. \tag{13}$$

Direct intra-group player communication is enabled in all transformer layers. We express the communication costs on a per-layer basis; the common number of layers does not affect the budget comparison. Let $B_b$ denote the number of allowed temporal frame pairs for direct intra-group player communication. The resulting preliminary directed player-pair budget is

$$B_{\text{pair}} = \left\lfloor \frac{\Delta\mathcal{C}}{B_bL^2} \right\rfloor. \tag{14}$$

The candidate group capacity is then bounded by

$$K = \min\left(N,\ \left\lfloor \frac{1}{2} + \sqrt{\frac{1}{4} + B_{\text{pair}}} \right\rfloor\right). \tag{15}$$

These preliminary limits restrict the candidate search. For the final check, $\mathcal{C}(\mathcal{G}^{(t)})$ counts the allowed directed query–key token pairs in the actual communication mask, excluding within-player attention shared by all methods. It accounts for overlapping memberships, global-only players, active and inactive hub attention, and unique direct player-to-player connections. We require

$$\mathcal{C}\big(\mathcal{G}^{(t)}\big) \le \mathcal{C}_{\text{central}}, \tag{16}$$

ensuring that the final hierarchical topology remains within the centralized communication budget. At each grouping step, the search is initialized with an empty set of local groups, under which all eligible players communicate through the global tokens. This topology is feasible under the centralized communication budget in our configuration. Candidate groups are then added only when the resulting partial topology satisfies all structural and communication constraints; infeasible candidates are discarded. At the first step, the relevance matrix is initialized from the camera state observed at $t = 0$, so no previous topology is required.

**Budget-Constrained Dynamic Grouping.** Given the relevance matrix $\mathbf{R}^{(t)}$ and the communication constraints $(M, K)$, we construct the active grouping topology $\mathcal{G}^{(t)}$. We first form candidate groups from high-relevance player relations and retain only candidates that satisfy the group-capacity constraint. Among feasible candidates, we select at most $M$ groups that prioritize player coverage while favoring high-relevance player relations and satisfying the communication budget.

Groups may overlap: each player belongs to at most two groups, and any two groups share at most one player. A shared player serves as an anchor, enabling indirect exchange between their local hubs through its visual representation. The notation $g(i)$ in the main text denotes one group containing player $i$; the same connection rule applies to each of its assigned groups. If unused group slots and budget remain, we attempt to cover unassigned players with additional feasible groups, reusing an assigned player as an anchor when needed and subject to the same overlap constraints. Players that remain uncovered communicate only through the global communication tokens.

Finally, the resulting variable-size groups are converted into a binary membership tensor

$$\mathbf{M} \in \{0,1\}^{f_z \times M \times N}, \tag{17}$$

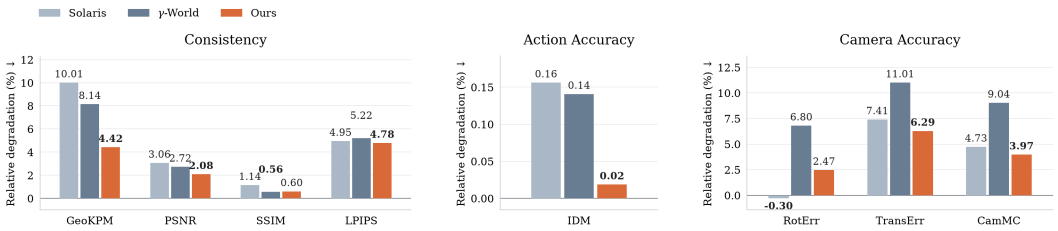



Figure 8: **Generalization from 4 to 6 players.** Relative performance degradation of models trained with 4 players compared with the same methods trained with 6 players, all evaluated with 6 players. Lower values indicate better preservation of performance; negative values indicate improvement. Bold labels mark the smallest degradation for each metric. Our method achieves the smallest degradation on six of the eight reported metrics.

Table 3: **Ablation on designs of group-token parameters.**

| Method | Consistency | | | | Action Accuracy | Camera Accuracy | | | Visual Quality | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GeoKPM↑ | PSNR↑ | SSIM↑ | LPIPS↓ | IDM↑ | RotErr↓ | TransErr↓ | CamMC↓ | FID↓ | FVD↓ |
| Copy group 0 | 45.01% | 16.43 | 0.4563 | 0.4797 | 94.08 | 4.413 | 12.91 | 16.01 | 26.43 | 199.0 |
| Copy group 1 | 43.63% | 16.41 | 0.4557 | 0.4799 | 94.04 | 4.606 | 13.15 | 16.38 | 26.39 | 200.2 |
| Average both | 43.96% | 16.41 | 0.4557 | 0.4802 | 94.04 | 4.765 | 13.32 | 16.62 | 26.82 | 202.5 |
| **Ours** | **54.33%** | **16.97** | **0.4659** | **0.4517** | **94.31** | **3.760** | **11.15** | **13.74** | **21.80** | **133.8** |

where $M_{t,g,i} = 1$ indicates that player $i$ is assigned to group slot $g$ at latent frame $t$. The membership count $\sum_g M_{t,g,i} \in \{0, 1, 2\}$ identifies global-only players, single-group members, and anchors, respectively. Each player connects to every assigned local hub; players sharing at least one group have direct attention connections. These connections form a binary mask and apply in all transformer layers together with the temporal attention constraints. Memberships may change at each grouping update. After converting the selected groups into the time-varying membership tensor, we additionally evaluate the exact clip-level communication cost induced by the complete attention mask and require it not to exceed $\mathcal{C}_{\text{central}}$.

# C Experimental Results

## C.1 Additional Results on Generalization to More Player Counts

To further quantify performance preservation, Figure 8 reports the relative degradation between $4 \rightarrow 6$ extrapolation and matched 6-player training. *Multics* exhibits the smallest degradation on most metrics, indicating stronger robustness to unseen player counts.

## C.2 Additional Ablation Study

**Effectiveness of Shared Group-token Parameters.** We train two sets of models with different designs for group-token parameters: (1) a model with non-shared group-token parameters, where each group is initialized with an independent set of learnable parameters; (2) our shared group-token model. Both models are trained on 4-player scenes with two groups and evaluated on 6-player scenes with three groups. For the unseen third group, we initialize the tokens of the model with non-shared group-token parameters by copying group 0, copying group 1, or averaging both. As shown in Table 3, our design of shared group-token parameters consistently outperforms all non-shared variants across all metrics. This demonstrates that sharing group-token parameters makes group slots exchangeable, allowing for varying numbers of groups and players, while maintaining robust multi-player generation.

## C.3 More Qualitative Results

Please see the additional qualitative results in the supplementary video, which has been compressed to 720p due to upload size limitations.

### C.4 Qualitative Results on Cross-Domain Applicability

We show qualitative results on EgoTeam (Peng et al., 2026) in Figure 9. EgoTeam contains multi-robot egocentric videos rendered in reconstructed real-world indoor scenes at 256×256 resolution and 5 FPS. Compared with $\gamma$-World, our method better preserves cross-view consistency and robustness across multiple agents, demonstrating that the proposed multi-player generation framework is applicable to multi-robot egocentric scenarios beyond gaming environments. The quantitative results in Table 1 also show that our method outperforms the baselines across multiple metrics.




Figure 9: **Qualitative comparison on EgoTeam.** We compare Solaris, $\gamma$-World, and *Multics* on multi-robot egocentric video generation.

## D Limitations

Our current work has several limitations that motivate future research:

1) We primarily focus on cross-player consistency, controllability, and robustness to cross-event interference. Maintaining persistent world states across multiple players over long horizons remains a challenging but important direction for future work.

2) *Multics* inherits limitations of the pre-trained video foundation model, including visual artifacts under large and dynamic human motions.