POPULATION-SCALABLE MULTI-AGENT WORLD MODELING ACROSS SCENE GEOMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

World models have recently achieved impressive progress in visual prediction and interactive generation, yet existing multi-agent systems are typically tied to predefined agent configurations and scene-specific environments. We propose **Khora**, a scalable multi-agent world model that supports both inference-time population expansion and generalization across geometrically distinct environments. Khora decouples shared world-state evolution from visual rendering: an action-conditioned state model maintains dynamic agents in a persistent Spatio-temporal Board (*STBoard*), while each observation is generated independently through a projection-based rendering interface. The renderer is trained with one target view per sample and reused across agents at inference without retraining. Crucially, scene geometry is provided to the renderer as an explicit mesh condition, allowing the same trained model to render observations that conform to different input environments without changing the underlying world dynamics or model architecture. This design enables Khora to scale along two complementary dimensions: many agents can interact within the same evolving world, while the same model can render observations across many different mesh-defined environments. Experiments demonstrate multi-agent consistency, action-conditioned interaction, generalization to unseen meshes, and scalability to large agent populations.

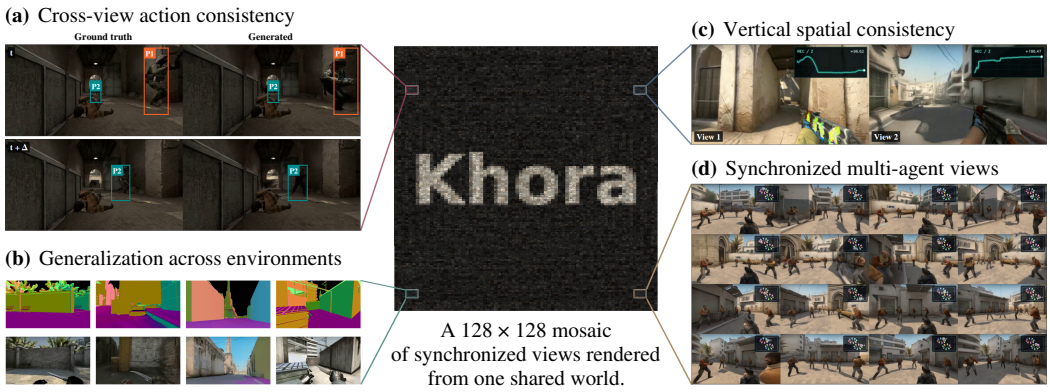

Figure 1: **Khora** renders observations from a shared world state. Its renderer, trained with one target view per sample, supports additional agents without retraining and follows mesh-defined scene layouts.

1 INTRODUCTION

Imagine a multiplayer battlefield where players act simultaneously and observe the same evolving world from different perspectives. A world model must do more than generate plausible egocentric videos: it must maintain a coherent, continuously evolving shared state consistent across agents, viewpoints, and actions. Recent advances in action-conditioned generation (Hafner et al., 2023; Wang et al., 2023; Bruce et al., 2024) and multi-agent modeling (Savva et al., 2026; Wu et al., 2026; Hu et al., 2026a) highlight these requirements. Yet a scalable simulator must allow agents to enter and leave at inference time without retraining or changing the model interface.

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

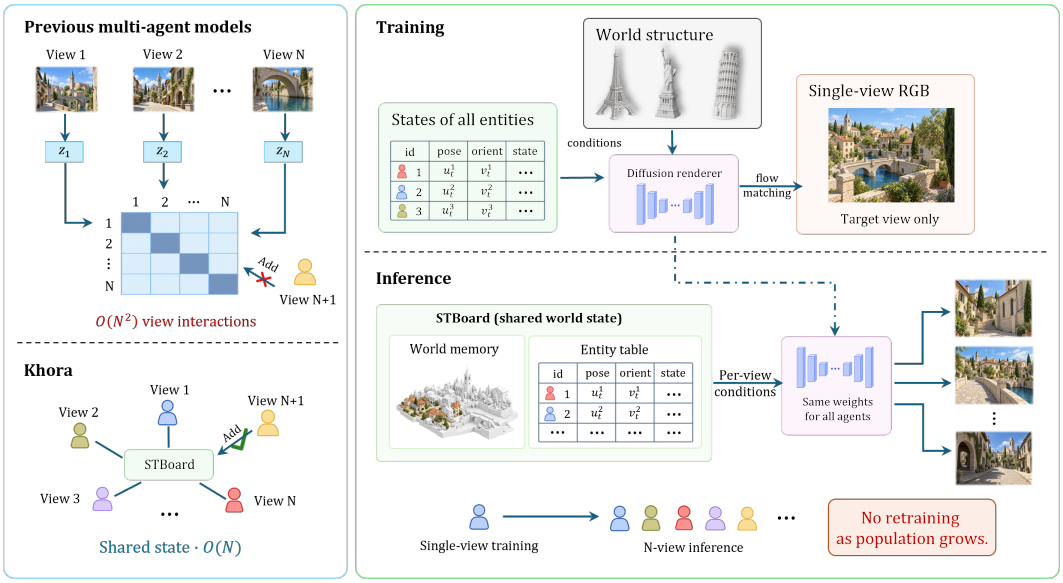

Figure 2: **Pairwise view coupling versus shared-state rendering.** Khora projects a persistent STBoard (Spatio-temporal Board) into view-specific conditions and reuses a shared renderer trained with one target view per sample for multi-view inference without retraining. New meshes provide geometry priors at inference.

Existing approaches enforce consistency through cross-agent interaction within video generators (Savva et al., 2026; Liu et al., 2026; Hu et al., 2026a;b) or shared representations and external memories (Po et al., 2026; Odyssey, 2026). Predefined view configurations can limit expansion, while direct view coupling places interaction costs inside the neural generator. **The governing rules of a physical world are population agnostic**: interactions depend on the entities present, but transition rules need not change with their number. We therefore treat population expansion as an inference-time operation on shared world state, rather than a change to agent-specific model components.

We propose **Khora**, a scalable multi-agent world model that decouples world-state evolution from visual rendering. Initialized from initial observations and a coarse scene prior, its action-conditioned model predicts world state and updates a persistent *STBoard* containing scene memory and dynamic entity states. Khora transforms relevant entities into the target camera coordinate frame and rasterizes them into a fixed-dimensional spatial condition. The renderer is trained with one target view per sample and reuses the same weights across views at inference; additional agents require only new state entries and rendering queries, without changing the renderer architecture or retraining.

This separation also supports mesh-guided scene expansion: a supplied mesh specifies scene geometry, allowing the same trained model to generate imagery following new spatial layouts. We demonstrate this capability with externally authored meshes, including white-box layouts constructed with language-model assistance.

Controlled qualitative evaluations and an interactive system prototype illustrate cross-view action consistency, persistence under partial occlusion, coherent rendering from synchronized viewpoints, and operation with agent populations not observed during training.

Our contributions are summarized as follows:

1. We formulate **inference-time population scalability** as an objective for multi-agent world models, requiring support for changing agent populations without architectural modification or retraining.
2. We introduce Khora, a scalable world model architecture that decouples world-state evolution from visual rendering. It combines a persistent shared STBoard, an action-conditioned pose and state transition model, and a projection-based per-view rendering interface.
3. We demonstrate that Khora directly supports multi-view generation at inference time through additional view queries, without architectural modification or retraining. It also maintains visual quality, multi-view consistency and action consistency. It achieves scalable real-time simulation with dominant rendering cost scaling linearly with the number of agents.

Table 1: **Comparison of representative multi-agent video world models.** Prior methods are ordered by their first arXiv submission date. ✓ indicates explicit support, △ partial or limited support, and – indicates that the capability is not explicitly established. The complexity shown for MIRA is inferred from full spatial attention over tiled views, not measured scaling beyond its four-player configuration.

Method	Persistent shared state	Dynamic population	Unseen-count generalization	Population-time scaling
Solaris (Savva et al., 2026)	–	–	–	$O(N^2)$
MultiGen (Po et al., 2026)	✓	△	–	$O(N)$
MultiWorld (Wu et al., 2026)	△	✓	△	$O(N^2)$
Gamma-World (Liu et al., 2026)	–	✓	△	$O(N)$
MIRA (Hu et al., 2026a)	△	–	–	$O(N^2)$
Khora (ours)	✓	✓	✓	$\approx \mathbf{O(N)}$

2 Related Works

World models and interactive generation. World models predict environment dynamics for planning and control (Ha and Schmidhuber, 2018; Hafner et al., 2023). Diffusion and transformer generators (Ho et al., 2020; Rombach et al., 2022; Dosovitskiy et al., 2021; Peebles and Xie, 2023) underpin action-conditioned video simulation in Genie, DIAMOND, AdaWorld, and Vid2World (Bruce et al., 2024; Alonso et al., 2024; Gao et al., 2025; Huang et al., 2025).

Multi-agent video world models. Solaris (Savva et al., 2026) couples two-player views through cross-view attention; MultiGen (Po et al., 2026) maintains external memory for editable multiplayer worlds. MultiWorld (Wu et al., 2026) uses a global state encoder, while Gamma-World (Liu et al., 2026) uses permutation-symmetric encodings and sparse hub attention. MIRA (Hu et al., 2026a) predicts four fixed-order views with tiled-latent attention and rolling history. MetaWorld (Hu et al., 2026b) derives multi-agent supervision from monocular videos and couples views through inter-branch attention. Khora trains its renderer with one target view per sample and supports population changes through fixed-dimensional queries of evolving shared state (Table 1).

Persistent scene memory. WorldMem (Xiao et al., 2025) stores frames with poses and timestamps. Context-as-Memory (Yu et al., 2025) retrieves frames by field-of-view overlap; VMem (Li et al., 2025) indexes views with surfels. Spatia (Zhao et al., 2025) updates a persistent point cloud through visual SLAM. WorldPlay (Sun et al., 2025) reconstructs historical context, while RELIC (Hong et al., 2025) retains compressed, camera-aware latent memory for real-time generation. Khora combines persistent appearance memory with an action-updated entity table and fixed-dimensional rendering queries, targeting inference-time population changes.

Geometry-guided image and scene generation. GQN (Eslami et al., 2018) separates scene representation from viewpoint-conditioned generation. ControlNet (Zhang et al., 2023) provides spatial conditioning, and SceneCraft (Yang et al., 2024b) projects semantic layouts into conditioning maps. GEN3C (Ren et al., 2025) renders a growing 3D cache into target-view conditions; WorldMesh (Schneider and Dai, 2026) and DAR (Chen et al., 2026) use mesh-derived conditions for image and video synthesis. Khora uses a fixed mesh scaffold with updated appearance and entity states to condition independently queried views; its unseen-mesh examples evaluate layout-conditioned generation.

Evaluation of visual world models. Image and video quality metrics (Wang et al., 2004; Zhang et al., 2018; Heusel et al., 2017; Unterthiner et al., 2018) do not directly establish multi-agent consistency. We complement them with human ratings, GT-referenced actor metrics, and geometric-and-color scene proxies related to MetaWorld (Hu et al., 2026b), distinguishing reference fidelity from agreement between generated views.

3 PRELIMINARIES

Consider a world with a variable set of agents I_t. At inference time, the model receives a coarse static world prior G, represented by a scene mesh, and a collection of initial observations $O_0 = \{o_0^i\}_{i \in I_0}$. At each subsequent time step t, the model receives a set of actions $A_t = \{a_t^i\}_{i \in I_t}$. The initial observations and static prior are used to initialize the world memory M and the initial agent poses:

$$M, \{p_0^i\}_{i \in I_0} = E(G, O_0). \tag{1}$$

During autoregressive rollout, no ground-truth future observations or poses are provided. The transition model updates shared state and predicts the next poses from current state and agent actions:

$$S_{t+1}, \{p_{t+1}^i\}_{i \in I_t} = F_\theta(S_t, \{p_t^i, a_t^i\}_{i \in I_t}; G). \tag{2}$$

Each requested observation is then independently decoded from the updated shared state:

$$o_{t+1}^i = R(p_{t+1}^i, S_{t+1}, G). \tag{3}$$

This formulation emphasizes three properties of a multi-agent world model. First, **persistence**: the shared state preserves scene and entity information across agents, viewpoints, and temporary occlusions. Second, **update**: agent actions modify both the predicted agent poses and the shared dynamic state. Third, **scalability**: observations are generated through independent view queries whose interface does not assume a predefined number of agents.

Our primary objective is inference-time population scalability: supporting changes in the active agent set without modifying or retraining the transition or rendering architecture. We additionally study mesh-guided scene expansion, in which the geometry prior G is replaced across rollouts by a new scene geometry G'. The model is then initialized for the new environment and generates observations conditioned on new structure. This setting evaluates whether the generated views follow the supplied geometry and remain consistent across views. Population scalability concerns changes in the entities inhabiting a world, whereas scene expansion concerns changes in the spatial structure of the world.

4 METHOD

Khora decouples world-state evolution from view synthesis (Figure 3). Initialized from observations and a coarse scene prior, the STBoard maintains scene geometry, appearance memory, and dynamic entities (Section 4.1). Actions update agent poses and entity states (Section 4.2); a shared renderer combines view-specific state queries with visual history, then back-projects generated appearance into scene memory (Section 4.3). Agent populations and requested views can change during inference. We use a coarse mesh to specify scene structure separately from appearance synthesis, a strategy also explored in mesh-conditioned generation and rendering (Schneider and Dai, 2026; Chen et al., 2026). Its geometry provides a common reference for view projection and geometric checks of proposed agent motion.

4.1 STBOARD

Multi-agent models may organize representations around predefined observers (Savva et al., 2026; Hu et al., 2026a), tying the rendering interface to the participating population.

Khora maintains a world-centric shared representation called the Spatio-temporal Board (*STBoard*). It associates a fixed scene geometry G with a persistent appearance memory M and a dynamic entity table D_t. The geometry provides the spatial scaffold of the environment. The appearance memory stores visual information associated with this scaffold, while the entity table represents active entities through attributes such as identity, pose, orientation, and task-dependent state.

The geometry remains fixed within a rollout, whereas the appearance memory and dynamic entity states are updated through different mechanisms. Agent actions drive updates to the entity table. Generated observations are back-projected onto the mesh to update the appearance memory, making accumulated scene appearance available to subsequent rendering queries. The initialization of scene memory and agent states uses the scene prior and available initial observations.

Each agent occupies an entity-table entry rather than a fixed view slot. Adding an agent requires a new entry and rendering queries, not a change to the renderer architecture.

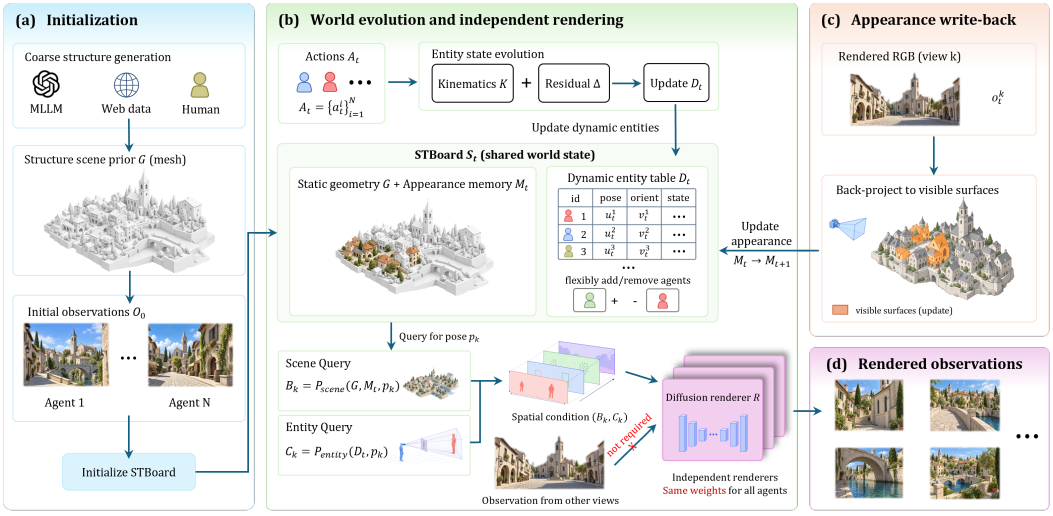

Figure 3: **Architecture of Khora.** Initial observations and a coarse scene prior initialize the shared STBoard (Spatio-temporal Board). Actions update the dynamic entity table through kinematic and residual transitions. Each requested view queries the shared state for pose- and scene-conditioned inputs, which are decoded by a shared-weight diffusion renderer. Generated RGB observations are back-projected onto visible surfaces to update the persistent appearance memory, while agents can be added or removed without retraining.

4.2 Action-conditioned World Evolution

Khora updates poses and shared state before rendering, so agent interactions are represented explicitly rather than inferred solely from image history.

For each agent, we form a deterministic proposal $\hat{p}_{t+1}^i = K(p_t^i, a_t^i)$ using the known action-to-motion mapping K. A learned residual corrects this proposal using the shared world state and static scene context:

$$p_{t+1}^i = \hat{p}_{t+1}^i + \Delta(S_t, G, \hat{p}_{t+1}^i, a_t^i). \tag{4}$$

The kinematic term captures control-induced translation and rotation, while the learned residual accounts for state-dependent effects such as collisions, terrain, and interactions with other agents. The refined poses and actions then update the shared dynamic state before rendering.

4.3 Geometry-conditioned Rendering and Appearance Update

Khora projects shared 3D agent states into the target camera view, providing explicit spatial conditions rather than requiring the renderer to infer multi-agent configurations from latent tokens.

View-conditioned spatial queries. For a target camera pose p_k, Khora queries the fixed geometry G and current appearance memory M_k for a scene condition B_k. Dynamic states at world time t are transformed into the same camera frame and rasterized into an entity condition C_k:

$$\begin{aligned} B_k &= \mathcal{P}_{\text{scene}}(G, M_k, p_k), \\ C_k &= \mathcal{P}_{\text{entity}}(S_t, p_k) \in \mathbb{R}^{H \times W \times d}. \end{aligned} \tag{5}$$

The entity condition has fixed dimensions as the population changes: additional entities modify its content without adding input slots or agent-specific renderer components.

Autoregressive latent generation. The renderer combines the spatial conditions with the target view's visual history. A causal Wan-VAE (Wan Team et al., 2025) encodes the initial observations into latent context. Conditioned on this context and the shared-world query, a diffusion transformer (Peebles and Xie, 2023) generates the next video latent block:

$$\hat{z}_k \sim p_\phi(z \mid H_k, B_k, C_k, p_k)\,, \tag{6}$$

where H_k is the latent history. The video decoder produces $\hat{o}_k = \mathcal{D}_{\text{VAE}}(\hat{z}_k)$, and each prediction updates the history for the next step (Appendix A, Figure 7); all views reuse the same renderer. The appendix also details windowed RGB decoding.

Shared memory across views. Latent history supplies per-view temporal context, while the entity table and appearance memory share agent states and scene information across views. Communication occurs through STBoard updates, not direct exchange between observation streams inside the neural renderer.

Computational cost. Let N_a denote the number of active agents and N_v the number of requested views. In view-coupled generators (Savva et al., 2026; Hu et al., 2026a), dense cross-agent communication inside the neural renderer can require $O(N_a N_v)$ expensive neural interactions, which becomes $O(N^2)$ when every agent receives a rendered view. Khora invokes the neural renderer independently for each requested view, resulting in $O(N_v)$ renderer evaluations. Constructing the conditioning maps still requires up to $O(N_a N_v)$ entity-to-view projections. However, the practical cost can be expressed as

$$T_{\text{Khora}} \approx N_v C_{\text{render}} + N_a N_v C_{\text{proj}}, \tag{7}$$

where $C_{\text{render}} \gg C_{\text{proj}}$. For up to 80 agents, projection and rasterization account for less than 8% of total inference latency.

4.4 Training Data and Inference Protocol

Khora is trained on Counter-Strike: Global Offensive (CS:GO) trajectories collected on a single map, Dust2, with aligned visual observations, actions, and structured state supervision.

The source episodes of the training set include up to 10 players, while each indexed training segment contains 2–4 player views. We sample exclusively from the valid temporal intervals specified in the training index. These intervals contain 44,486,627 player-view frames at 32 FPS, totaling 411.9 hours of aggregate player-view video.

A coarse world prior (mesh) initializes the shared scene representation. At inference time, Khora receives this prior, initial observations, and subsequent actions, and autoregressively generates future poses, dynamic states, and observations.

5 Experiments

We evaluate visual quality, multi-agent consistency, computational scaling, dynamic populations, and mesh-guided generalization. External methods are ordered by first public release, followed by the Khora ablation and full model.

Khora w/o Cross-Agent State masks other-agent state conditions at inference while retaining scene memory, ego conditions, and visual history. It uses the same renderer checkpoint and inference settings without retraining; visual history may still contain other agents.

5.1 Visual Quality

We assess visual fidelity using paired metrics (PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018)) and distribution-level metrics (FID (Heusel et al., 2017) over individual frames and FVD (Unterthiner et al., 2018) over complete generated clips).

Paired metrics compare generated and synchronized GT frames, averaging over agents and timesteps. Scores use a common resolution but retain model-specific temporal windows. Evaluation splits, sample counts, and scoring windows are detailed in Appendix B.

Table 2 reports per-view quality across population sizes. Khora improves over its input ablation on most metrics while retaining a population-independent rendering interface. Appendix B details checkpoints, preprocessing, and evaluation scope.

5.2 Multi-Agent Consistency

We evaluate multi-agent consistency using human ratings on a five-point scale and scene-level point-cloud geometry and color scores inspired by MetaWorld's CVC formulation (Hu et al., 2026b). Specifically, **CVC-XView** evaluates the similarity of reconstructed generated views; **CVC-GT**

Table 2: **Quantitative comparison of visual quality.** Scores use a common resolution; temporal windows and available evaluation cohorts differ (Appendix B).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
2 Views (10,000 cases / 20,000 videos)					
Solaris (Savva et al., 2026)	20.1465	0.5273	0.1835	**4.3726**	**11.4263**
Gamma-World (Liu et al., 2026)	15.5611	0.4332	0.5607	53.9303	297.1115
MIRA (Hu et al., 2026a)	21.7290	0.5973	0.3927	16.0886	231.9198
Khora w/o Cross-Agent State	26.0427	0.7125	0.1826	8.7262	51.6657
Khora	**26.2425**	**0.7130**	**0.1789**	7.7098	36.5529
4 Views (10,000 cases / 40,000 videos) *No comparable results for Solaris (configuration limits).*					
Gamma-World (Liu et al., 2026)	16.9473	0.5025	0.4856	55.8759	272.2558
MIRA (Hu et al., 2026a)	20.8771	0.6072	0.3570	16.3253	198.0576
Khora w/o Cross-Agent State	24.9196	0.7017	0.1851	11.3010	73.4917
Khora	**25.2825**	**0.7020**	**0.1772**	**7.9821**	**38.6849**
8 Views (159 cases / 1,272 videos) *No comparable results for Solaris, Gamma-World, or MIRA (configuration limits).*					
Khora w/o Cross-Agent State	26.2282	0.7573	0.1544	30.1706	104.9082
Khora	**26.5775**	**0.7573**	**0.1477**	**27.0311**	**53.6148**

Note. All external baselines completed all required training stages on our dataset. Checkpoints, evaluation settings, and sample counts are detailed in Appendix B.1.

Table 3: **Human evaluation and automatic world consistency.** Human ratings use a five-point scale; CVC scores lie in $[0, 1]$. Higher is better. CVC-XView is evaluated on 10,000 common cases across all five models. CVC-GT and CVC-XView use different evaluation cohorts and temporal windows; Appendix D details both protocols and their component scores.

	Human evaluation			Automatic world scores	
Method	**Action ↑**	**World ↑**	**Overall ↑**	**CVC-GT ↑**	**CVC-XView ↑**
Solaris (Savva et al., 2026)	3.000	4.105	3.553	0.918310	0.836662
Gamma-World (Liu et al., 2026)	2.027	2.786	2.406	0.897661	0.836490
MIRA (Hu et al., 2026a)	3.170	4.009	3.589	0.922486	0.830889
Khora w/o Cross-Agent State	2.075	4.065	3.070	0.931900	0.838207
Khora	**3.570**	**4.213**	**3.892**	**0.932200**	**0.838773**

compares each generated view with its GT, then averages over views. These proxies do not establish actor synchronization. All five methods use the same human-rating protocol (Appendix C.1); the evaluation sets and temporal windows for the automatic metrics are detailed in Appendix D.

Khora receives the highest human ratings (Table 3). Its small CVC-XView gain over the ablation is descriptive, not evidence of synchronized character motion.

GT-referenced cross-view action consistency. We measure the other player's location, image motion, pose, and appearance against GT in one ego view per case. GT-only eligibility checks select 1,643 of 10,000 held-out evaluation cases. Khora uses supplied trajectories and state conditions: this evaluates state-conditioned rendering, not autonomous dynamics or direct agreement between generated views (Appendix E.1).

Table 4 separates two physical windows because Solaris and MIRA have no common purely generated interval. With checkpoint and inference settings fixed, removing other-agent conditions degrades localization, motion, pose, and appearance in both windows. This supports their contribution to visible actor fidelity, not long-horizon dynamics. Full metrics and paired confidence intervals appear in Appendices E.3 and E.4.

5.3 Scalability and Real-Time Performance

Figure 4 reports measured scaling from 1–80 agents: latency rises from 107.16 ms to 116.73 ms per rollout step and per-view FPS falls from 37.33 to 34.27, with view synthesis dominant. Peak

Table 4: **Cross-view action consistency with a GT reference.** Each panel scores the same 1,643 eligible cases at matched physical frames. Conditioning histories differ across model families; Khora and its ablation share the checkpoint and inference settings. PCK uses a threshold of 20% of GT box height; FG denotes fixed GT foreground. MAE is lower-is-better.

Method	F1 ↑	Recall ↑	Direction ↑	PCK ↑	FG-SSIM ↑	FG-MAE ↓
Relative frames 13–32 (zero-based)						
Solaris (Savva et al., 2026)	0.1683	0.1611	0.5267	0.2111	0.1803	0.1647
Gamma-World (Liu et al., 2026)	0.0619	0.0525	0.1963	0.0500	0.1568	0.2000
Khora w/o Cross-Agent State	0.2842	0.1850	0.0654	0.1723	0.2487	0.1417
Khora	**0.9041**	**0.8721**	**0.7121**	**0.7438**	**0.3910**	**0.0834**
Relative frames 38–57 (zero-based)						
Gamma-World (Liu et al., 2026)	0.0195	0.0151	0.0921	0.0181	0.1476	0.2261
MIRA (Hu et al., 2026a)	0.4525	0.4299	0.4507	0.4129	0.2555	0.1268
Khora w/o Cross-Agent State	0.0087	0.0047	0.0043	0.0045	0.2201	0.1625
Khora	**0.9024**	**0.8645**	**0.6929**	**0.7385**	**0.3609**	**0.0868**

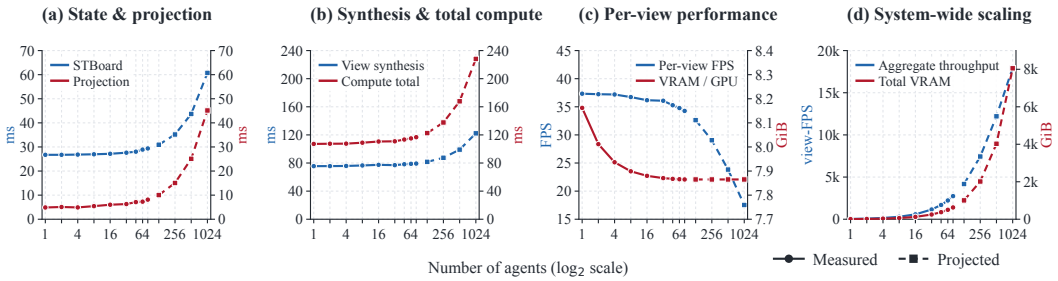

Figure 4: **Runtime scalability with increasing agent population.** Blue and red denote the left- and right-axis metrics in each panel, respectively. Solid circles are measurements (1–80 agents); dashed squares are projections (128–1024 agents). Curves are drawn separately across the measured/projected boundary. Aggregate throughput equals agent count times per-view FPS; each requested view is synthesized on its respective GPU.

per-GPU VRAM is nearly constant; distributed aggregate throughput rises from 37.3 to 2741.3 view-fps. Projections beyond 80 agents are extrapolations (Appendix F).

5.4 Inference-Time Dynamic Population

We finally evaluate whether the active population can change within a single autoregressive rollout. The sequence begins with two agents. Additional agents are introduced at later timesteps, increasing the population from two to four and then from four to eight. Several agents subsequently leave the environment, reducing the population again. Newly introduced agents are initialized from their first observations and appended to the dynamic entity table, while departing agents are removed from the active set. The transition model, STBoard interface, and neural renderer remain unchanged throughout the rollout.

Figure 5 illustrates that Khora can accommodate population changes during an ongoing simulation. Existing agents preserve their trajectories and observations when new agents are introduced, while the new agents become visible from relevant viewpoints after being inserted into the shared state. This experiment demonstrates that population scalability is an inference-time capability rather than a fixed configuration determined during training.

5.5 Mesh-guided Generalization to Unseen Environments

We examine whether Khora can generate scene imagery conditioned on geometry not encountered during training. We supply previously unseen meshes as structural priors and use the same trained model weights as in the main experiments, without additional weight updates. Figure 6 compares the input mesh structures with the corresponding generated observations from matched camera

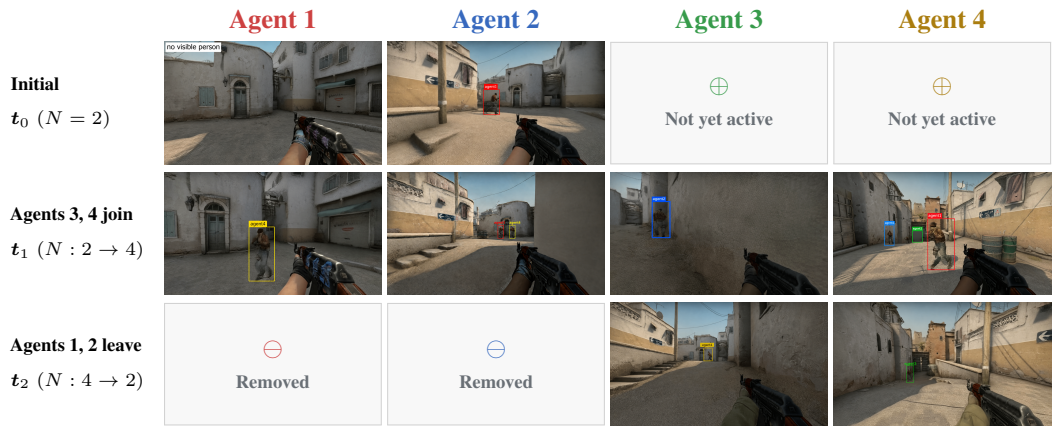

Figure 5: **Dynamic population within a single rollout.** Agents enter and leave while the shared world persists; model weights and the transition and rendering architecture remain unchanged.

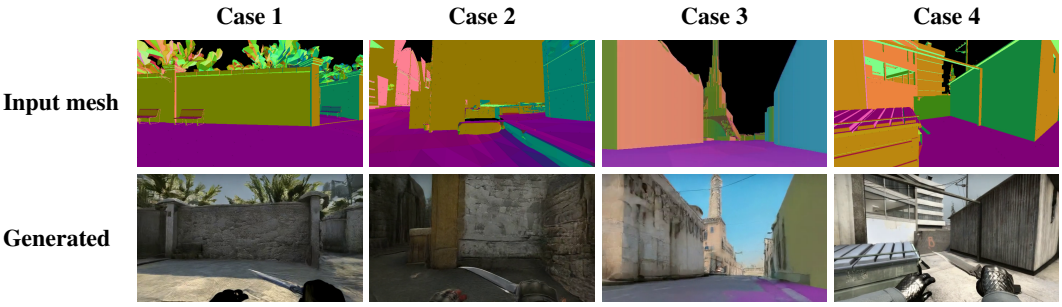

Figure 6: **Mesh-guided generation in unseen environments.** Each column pairs an unseen mesh (top) with the corresponding generated observation (bottom). The same trained weights are used without additional updates; the generated views follow the supplied scene geometry.

viewpoints. These examples show that Khora synthesizes visual appearance that follows the supplied spatial layouts, demonstrating mesh-guided scene generalization beyond the training environments.

6 Discussion & Limitations

Khora requires authored geometry, reliable memory, and rendering resources. Mesh-guided results do not establish dynamics generalization; motion, collisions, action consequences, structural accuracy, and long-horizon consistency require broader evaluation. Appearance back-projection can contaminate memory with generation or alignment errors under occlusion or conflicting observations. Rendering costs grow with views; projection costs up to $O(N_a N_v)$, with latency depending on population, scene representation, scheduling, and hardware.

Future work includes automatic geometry acquisition, closed-loop evaluation, confidence-aware updates, cross-view conflict handling, visibility-aware projection, and efficient rendering.

7 Conclusion

We present Khora, a multi-agent world model that separates shared-state evolution from per-view rendering through a persistent STBoard, enabling dynamic populations without architectural changes or retraining. Experiments assess visual quality, multi-agent consistency, population flexibility, and computational scaling, while qualitative results demonstrate generation conditioned on unseen meshes with unchanged model weights. We hope this work provides a practical step toward scalable open-world simulators capable of supporting large numbers of interacting agents, and encourages future research on world-centric representations and scalable interactive world models.

8 LLM USAGE

Large Language Models (LLMs) were used as assistive tools during the preparation of this work. In particular, we used LLMs to support manuscript writing and polishing, including language refinement, grammar checking, sentence rephrasing, and improving the readability and clarity of the text. LLMs were also used to assist with certain aspects of experimental implementation and execution, such as code development and debugging, scripting, and troubleshooting during experiments.

LLMs were not used to independently formulate the research questions, develop the core research ideas, design the methodology, or determine the experimental design. All scientific decisions, including the formulation of the proposed approach, the design of experiments, and the interpretation of experimental results, were made by the authors. Outputs produced with LLM assistance were reviewed, verified, and, where necessary, modified by the authors before being incorporated into the research or manuscript.

The authors take full responsibility for the content of the manuscript and the validity of the reported results. LLMs were used solely as assistive tools and did not act as autonomous contributors to the scientific work.

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, 2024. URL https://arxiv.org/abs/2405.12399.

Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

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. *arXiv preprint arXiv:2402.15391*, 2024.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the Kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Junhao Chen, Mingjin Chen, Henghaofan Zhang, Minglin Chen, Liaoyuan Fan, Boran Zhang, Saining Zhang, Mingze Sun, Hao Zhao, Ruqi Huang, Zhihao Li, and Yufei Wang. Video models as native 4d renderers: World-grounded conditioning from animated mesh, 2026. URL https://arxiv.org/abs/2608.00094.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018. doi: 10.1126/science.aar6170.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions, 2025. URL https://arxiv.org/abs/2503.18938.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

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, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Amélie Royer, Manu Orsini, Alyx Liao, 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. Multiplayer interactive world models with representation autoencoders. *arXiv preprint arXiv:2607.05352*, 2026a. URL https://arxiv.org/abs/2607.05352.

Teng Hu, Mingchun Lu, Yating Wang, Jiangning Zhang, Jinkun Hao, Ye Pan, Ran Yi, Lizhuang Ma, and Dacheng Tao. MetaWorld: Scaling multi-agent video world model from single-view video data. *arXiv preprint arXiv:2606.02753*, 2026b.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models, 2025. URL https://arxiv.org/abs/2505.14357.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO. Software, 2023. URL https://github.com/ultralytics/ultralytics. Evaluation uses version 8.3.203 with YOLOv8 weights.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25690–25699, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014.

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*, 2026. URL https://arxiv.org/abs/2605.28816.

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in PyTorch, 2020. URL https://github.com/toshas/torch-fidelity.

Odyssey. Agora-1: The multi-agent world model. https://odyssey.ml/introducing-agora-1, 2026. Accessed 2026-07-09.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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. URL https://arxiv.org/abs/2603.06679.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

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. URL https://arxiv.org/abs/2602.22208.

Manuel-Andreas Schneider and Angela Dai. Worldmesh: Generating navigable multi-room 3d scenes via mesh-conditioned image diffusion, 2026. URL https://arxiv.org/abs/2603.22972.

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 modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2018. URL https://arxiv.org/abs/1812.01717.

Wan Team, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving, 2023. URL https://arxiv.org/abs/2309.09777.

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.

Haoyu Wu, Jiwen Yu, Yingtian Zou, and Xihui Liu. MultiWorld: Scalable multi-agent multi-view video world models. *arXiv preprint arXiv:2604.18564*, 2026. URL https://arxiv.org/abs/2604.18564.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *arXiv preprint arXiv:2406.09414*, 2024a. URL https://arxiv.org/abs/2406.09414.

Xiuyu Yang, Yunze Man, Jun-Kun Chen, and Yu-Xiong Wang. SceneCraft: Layout-guided 3d scene generation. In *Advances in Neural Information Processing Systems*, 2024b. URL https://arxiv.org/abs/2410.09049.

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. *arXiv preprint arXiv:2506.03141*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. URL https://arxiv.org/abs/1801.03924.

Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory. *arXiv preprint arXiv:2512.15716*, 2025.

APPENDIX GUIDE

The appendix details the renderer, evaluation protocols, and additional results.

Section	Topic	Page
A	Design of diffusion renderer	14
B	Experimental setup and visual quality	16
C	Human evaluation	18
D	Cross-view world consistency	20
E	Cross-view action consistency	23
F	Runtime and scalability	39
G	Additional qualitative results	41

Arrows indicate preferred metric directions. Bold and underlined values mark the best and second-best results within each comparison group; two-method comparisons highlight only the better value. Counts, uncertainty statistics, and diagnostic controls are not ranked.

A DESIGN OF DIFFUSION RENDERER

Khora renders each requested view using a conditional latent diffusion transformer (Peebles and Xie, 2023) and a frozen 3D video VAE (Wan Team et al., 2025). Figure 7 illustrates the autoregressive process. The renderer maintains a sliding history of three video latents, predicts one additional latent per generation step, and decodes the updated latent window into RGB observations. All views share the same renderer parameters, while maintaining their own temporal histories.

Latent representation and initialization. Let $\mathcal{E}$ and $\mathcal{D}$ denote the frozen 3D-VAE encoder and decoder, respectively. The VAE has a temporal compression factor of four: each newly predicted latent frame corresponds to four RGB frames. The initial RGB context is encoded into video latents, and the first prediction uses the three-latent history

$$H_2 = [z_0, z_1, z_2]. \tag{8}$$

Each z_j denotes a temporal latent frame containing spatial features, rather than an individual spatial token. Both the encoder and decoder remain frozen during renderer training.

Shared-world conditioning. At each generation step, the DiT receives the latent history together with the target-view condition obtained from the shared world representation. We denote this condition by c_{n+1}; it comprises the scene and entity conditions associated with the target camera query, as described in the main method.

The scene condition conveys geometry and available appearance memory, while the entity condition describes the projected dynamic state. The history supplies temporal visual context, and the shared-world condition specifies the scene configuration to be rendered.

Next-latent prediction. The renderer generates one new temporal latent conditioned on the preceding three latents and the target-view condition:

$$\hat{z}_{n+1} \sim p_\phi(z_{n+1} \mid H_n, c_{n+1}), \qquad n \geq 2, \tag{9}$$

where p_ϕ is implemented by the conditional latent DiT.

For compact notation, define

$$\bar{z}_j = \begin{cases} z_j, & j \in \{0, 1, 2\}, \\ \hat{z}_j, & j \geq 3. \end{cases} \tag{10}$$

The history at step n is then

$$H_n = [\bar{z}_{n-2}, \bar{z}_{n-1}, \bar{z}_n]. \tag{11}$$

Thus, the first prediction is conditioned on encoded observations, whereas subsequent predictions progressively depend on generated latents.

Sliding-window history update. After $\hat{z}_{n+1}$ is predicted, the oldest latent is removed and the new latent is appended:

$$H_{n+1} = [\bar{z}_{n-1}, \bar{z}_n, \hat{z}_{n+1}]. \tag{12}$$

The first three updates are

$$[z_0, z_1, z_2] \longrightarrow \hat{z}_3, \tag{13}$$
$$[z_1, z_2, \hat{z}_3] \longrightarrow \hat{z}_4, \tag{14}$$
$$[z_2, \hat{z}_3, \hat{z}_4] \longrightarrow \hat{z}_5. \tag{15}$$

After these updates, the history consists entirely of generated latents. No future ground-truth RGB observations are introduced during rollout. Predicted latents are passed directly to the next history window; decoding and re-encoding the generated RGB output are not required for this history update.

Windowed RGB decoding. After a new latent frame is predicted, the frozen decoder processes the updated three-latent window:

$$\widetilde{O}_{n+1} = \mathcal{D}([\bar{z}_{n-1}, \bar{z}_n, \hat{z}_{n+1}]). \tag{16}$$

Only the four RGB frames corresponding to the final latent frame $\hat{z}_{n+1}$ are retained:

$$\hat{O}_{n+1} = \mathrm{Tail}_4\Big(\widetilde{O}_{n+1}\Big), \tag{17}$$

where Tail_4 selects the last four decoded frames. The preceding two latent frames provide temporal context for decoding, and their corresponding RGB outputs are discarded. Each autoregressive step therefore appends four new frames to the output sequence, without duplicating frames from overlapping decoding windows. The predicted latent is also retained directly in the history for the next generation step.

Temporal history and shared appearance memory. The latent history and the STBoard appearance memory serve different purposes. The three-latent history is local to an observation stream and supplies recent temporal context. The appearance memory is associated with the shared scene geometry and can retain information beyond this local window. Decoded observations are back-projected onto the mesh to update the appearance memory, as described in the main method. Subsequent view queries can access the updated scene information through their conditions. This appearance update is distinct from the direct latent-history update.

Extension to multiple views. For multiple requested views, the same autoregressive procedure is applied with a separate history H_n^i and target-view condition c_{n+1}^i for each view i:

$$\hat{z}_{n+1}^i \sim p_\phi\big(z_{n+1}^i \mid H_n^i, c_{n+1}^i\big). \tag{18}$$

The renderer parameters and three-latent history length do not change with the number of views. Additional views require additional histories and rendering queries, while scene and entity information are obtained from the shared STBoard. The fixed history bounds the temporal context processed by each rendering query; it does not impose a constant memory cost on the complete multi-view system.

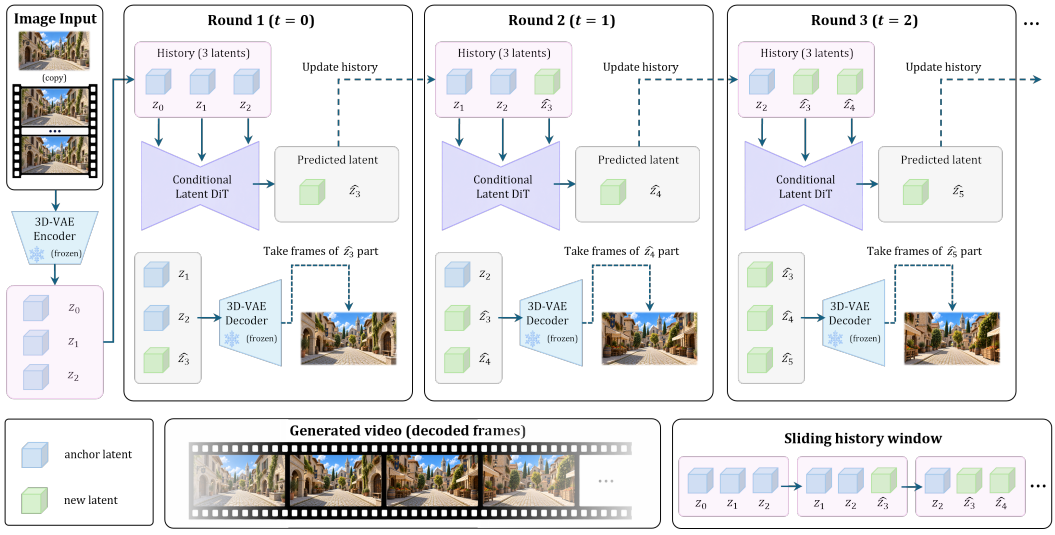

Figure 7: **Autoregressive latent rendering in Khora.** A frozen 3D-VAE encoder maps the initial RGB context into video latents. At each generation step, a conditional latent DiT predicts one new latent from a history of three latents. The prediction is appended to the history after removing its oldest element. The frozen 3D-VAE decoder processes the updated three-latent window. Only the four RGB frames corresponding to the newly predicted latent frame are retained. Blue blocks denote latents encoded from the initial observations, and green blocks denote generated latents. The same DiT and VAE parameters are reused throughout the rollout.

B Experimental Setup and Visual Quality

B.1 Baseline Checkpoints and Preprocessing

For Solaris (Savva et al., 2026), the CS:GO adaptation uses the Dust2 Stage 2 checkpoint at training iteration 120,000. Its multiplayer configuration jointly generates two synchronized views with 50 denoising steps. Each view uses one RGB conditioning frame and 32 subsequent generated frames; the checkpoint is evaluated only in the two-view setting. Input observations are resized with aspect-ratio-preserving letterboxing to 360×640. Saved predictions have resolution 352×640, and the common-window visual-quality scorer resizes them to 360×640 using bilinear interpolation. Dimensions here are height × width.

For Solaris, the conditioning observation is physical frame 0 and the first generated frame corresponds to physical frame 1. Its native evaluation covers physical frames 1–32; the common-window evaluation uses frames 13–32 on 10,000 two-view cases.

For Gamma-World (Liu et al., 2026), we use the Stage 2 checkpoint at training iteration 10,000. Its native 320×480 output is resized to the common 360×640 scoring resolution. The reported values are computed on the complete 10,000-case two-view evaluation set over frames 38–79.

For the four-view Gamma-World evaluation, we use the same Stage 2 checkpoint with four agent slots. We evaluate 10,000 cases (40,000 videos), scoring frames 38–79 after one RGB conditioning frame per view. GT preprocessing follows the four-view pipeline: resize to 288×512, normalize and quantize through bfloat16, then bilinearly resize to 360×640.

For MIRA (Hu et al., 2026a), we use the two-view checkpoint at training step 15,000 for the two-view baseline and the corresponding four-view checkpoint at training step 3,000 for the four-view baseline. Both evaluations use frames 38–79 after conditioning on 38 RGB frames. MIRA's native 288×512 decoded output is resized to the common 360×640 scoring resolution, and the paired ground truth is resized from 480×832. MIRA does not provide an 8-view result in this experiment.

Table 5: **Detailed Khora visual-quality results.** Khora and Khora w/o Cross-Agent State share their view-specific cohort. PSNR is in dB; six decimals are shown. The 8-view cohort is substantially smaller.

Views	Model	Cases	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
2	Khora w/o Cross-Agent State	10000	26.042689	0.712535	0.182607	8.726227	51.665660
2	Khora	10000	**26.242469**	**0.713042**	**0.178877**	**7.709788**	**36.552903**
4	Khora w/o Cross-Agent State	10000	24.919565	0.701700	0.185143	11.301036	73.491736
4	Khora	10000	**25.282468**	**0.702034**	**0.177186**	**7.982100**	**38.684905**
8	Khora w/o Cross-Agent State	159	26.228233	0.757253	0.154418	30.170570	104.908220
8	Khora	159	**26.577453**	**0.757305**	**0.147736**	**27.031112**	**53.614786**

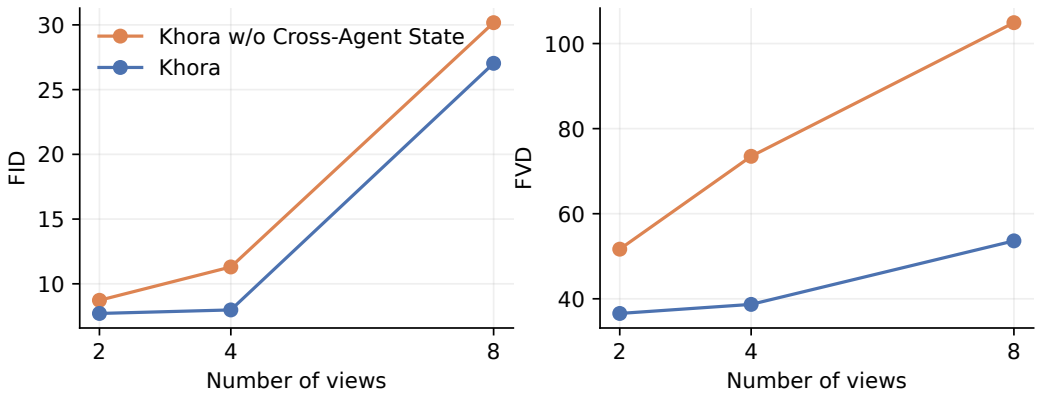

Figure 8: **Visual quality across the available view settings.** The six points in each panel come directly from Table 5. Lower is better. Lines connect separate cohorts and are descriptive; particularly, the eight-view cohort has only 159 cases. No uncertainty bars are inferred from aggregate FID/FVD values.

B.2 Visual Quality: Evaluation Protocol and Results

Table 5 reports the detailed Khora and input-ablation results underlying Table 2. The two- and four-view settings each contain 10,000 cases (20,000 and 40,000 videos); the eight-view setting contains 159 cases (1,272 videos). All six rows score the 68-frame suffix, source frames 13–80. Evaluation samples are held out from model training. Khora scores frames 13–80 after 13 RGB context frames, while MIRA scores frames 38–79 after 38 context frames. Gamma-World also scores frames 38–79, with one RGB context frame. Scoring windows therefore differ across models.

Feature and frame conventions. PSNR averages frame-level RGB scores in dB; SSIM (Wang et al., 2004) uses an 11×11 Gaussian kernel with $\sigma = 1.5$; LPIPS v0.1 (Zhang et al., 2018) uses AlexNet (Krizhevsky et al., 2012). FID (Heusel et al., 2017) uses 2,048-dimensional torch-fidelity (Obukhov et al., 2020) Inception (Szegedy et al., 2016) features and one deterministically selected frame per video. Its local suffix index is $(1000003s + 10007v) \bmod 68$, with sample index s and view index v. FVD (Unterthiner et al., 2018) uses the 400-dimensional output of the TF-Hub Kinetics-400 I3D model (Carreira and Zisserman, 2017), with local indices $[0, 4, 9, 13, 18, 22, 27, 31, 36, 40, 45, 49, 54, 58, 63, 67]$. This feature definition is part of the metric: values from other I3D representations cannot be compared as if they used the same evaluator.

Detailed baseline results. Table 6 reports the Gamma-World and MIRA visual-quality results to six decimal places. These score 42 frames, 38–79, at 640×360, unlike Khora's 68-frame suffix. Gamma-World conditions on one RGB frame and MIRA on 38. Their FID selects one frame per

Table 6: **Detailed baseline visual-quality results.** All rows score source frames 38–79 on 10,000 cases (20,000 videos for two views; 40,000 for four views). PSNR is in dB. These horizons differ from the Khora rows above.

Method	Views	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
Gamma-World (Liu et al., 2026)	2	15.561124	0.433187	0.560724	53.930308	297.111511
Gamma-World (Liu et al., 2026)	4	16.947278	0.502547	0.485552	55.875949	272.255839
MIRA (Hu et al., 2026a)	2	21.728992	0.597260	0.392728	16.088566	231.919817
MIRA (Hu et al., 2026a)	4	20.877136	0.607222	0.356992	16.325321	198.057638

video with the analogous deterministic rule modulo 42. MIRA's two-view evaluation directly resizes decoded GT frames to 640×360 using bilinear interpolation. Its four-view evaluation first resizes GT frames to 832×480, normalizes them to $[-1, 1]$ in bfloat16, and then bilinearly resizes them to 640×360 before rounding to 8-bit RGB. These preprocessing differences should be considered when comparing the two settings.

C Human Evaluation

C.1 User-Study Protocol

The rating interfaces present synchronized multi-view clips in randomized, model-anonymized order. Participants rate each assigned sample on a five-point scale, and scores are averaged over the ratings in the cohorts specified below. All five methods follow the same rating protocol, including the five-point scale, model-anonymized presentation, and scoring criteria, providing a fair comparison under shared evaluation criteria. Participant and case counts for each evaluation cohort are reported below. The two criteria are defined as follows:

Cross-view action consistency. Raters judge whether an action performed by one agent produces temporally and spatially compatible consequences in the other observations. They consider actor motion, interaction events, and event timing across synchronized views. Low scores correspond to contradictions such as an entity moving in one view but remaining stationary in another, or the same event occurring at different times.

Cross-view world consistency. Raters judge whether shared visible regions agree on static scene geometry, fixed objects, materials, and spatial relations. Normal differences in perspective, occlusion, lighting, and field of view are not penalized. Dynamic agents and their actions are assessed under action consistency instead.

Paired Khora study. For Khora and its without-cross-agent-state ablation, each participant received 150 randomized, model-blinded samples: 25 source cases for each combination of two model variants and three view counts (2, 4, and 8). Within each view count, the same 25 source cases were assigned for both variants. The interface synchronized playback, seeking, and playback speed across the generated views; raters compared these views with one another, rather than against a GT video.

C.2 Complete Ratings and Distributions

Table 7 reports the human-rating scores summarized in Table 3. Rating distributions, participant and sample counts, and aggregation conventions are detailed below. External baselines follow first-public-release order.

Human-rating distributions. The baseline evaluation includes 338 submitted ratings from 12 participants: 114 for Solaris and 112 each for Gamma-World and MIRA, covering ten source cases per model. Table 8 and Figure 9 include these records together with 600 ratings each for Khora and its ablation, whose participant selection and view-count breakdown are described below. Only submitted ratings are included. These are repeated observations within participants and cases; an ordinary independent-sample standard error would be inappropriate.

Table 7: **Human consistency ratings matching the main table.** Scores use a five-point scale; higher is better.

Method	Action ↑	World ↑	Overall ↑
Solaris (Savva et al., 2026)	3.000	4.105	3.553
Gamma-World (Liu et al., 2026)	2.027	2.786	2.406
MIRA (Hu et al., 2026a)	3.170	4.009	3.589
Khora w/o Cross-Agent State	2.075	4.065	3.070
Khora	**3.570**	**4.213**	**3.892**

Table 8: **Human-rating distributions for all five models.** Baselines have 12 raters and 10 source cases each. For Khora and its ablation, eight participants completed all assignments, yielding 600 ratings per model, pooled equally over 2, 4, and 8 views. Columns 1–5 give exact counts; SD is descriptive. Mean ranks compare methods within each criterion.

Method	Criterion	n	Mean ↑	SD	1	2	3	4	5
Solaris (Savva et al., 2026)	Action	114	3.000000	1.240896	12	33	30	21	18
Solaris (Savva et al., 2026)	World	114	4.105263	1.162606	5	8	17	24	60
Gamma-World (Liu et al., 2026)	Action	112	2.026786	1.142621	47	35	14	12	4
Gamma-World (Liu et al., 2026)	World	112	2.785714	1.188750	18	29	34	21	10
MIRA (Hu et al., 2026a)	Action	112	3.169643	1.192230	7	30	32	23	20
MIRA (Hu et al., 2026a)	World	112	4.008929	0.925081	2	5	20	48	37
Khora w/o Cross-Agent State	Action	600	2.075000	1.192369	253	171	84	62	30
Khora w/o Cross-Agent State	World	600	4.065000	0.899327	9	27	89	266	209
Khora	Action	600	**3.570000**	1.103235	21	93	144	207	135
Khora	World	600	**4.213333**	0.901393	7	23	82	211	277

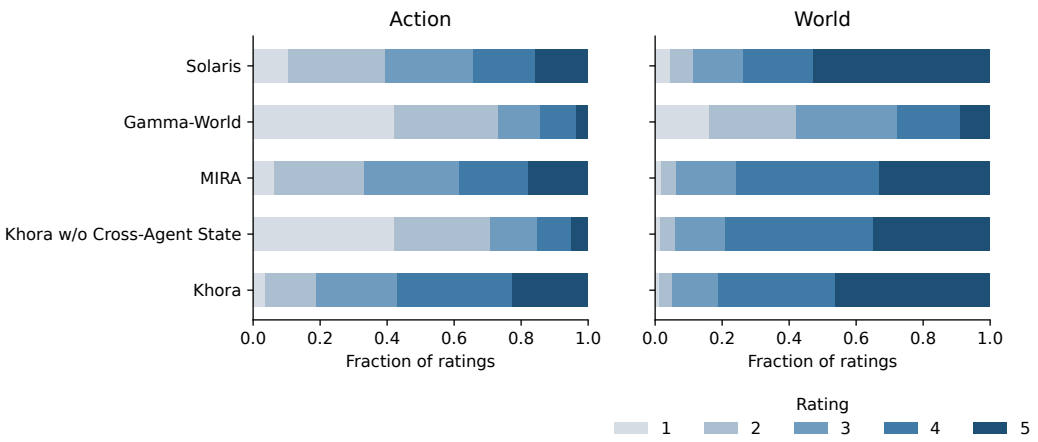

Figure 9: **Distribution of available human ratings.** Each bar sums to one within a model and criterion. Darker segments denote higher ratings on the five-point scale. Counts and descriptive means are given in Table 8; the Khora view-count breakdown appears in Table 9. All methods use the same five-point rating protocol. Solaris, Gamma-World, and MIRA have 114, 112, and 112 ratings from 12 participants; Khora and its ablation each have 600 ratings from eight participants who completed all assignments. The panels do not imply independent ratings or statistical significance.

Table 8 reports criterion means and standard deviations to six decimal places. The main table rounds each criterion mean to three decimals and computes Overall from the unrounded criterion means. Khora and ablation ratings and sample counts are further broken down by view count in Section C.3.

C.3 Khora and Ablation: Paired Human Evaluation

Participants and evaluation cohort. Fourteen participants submitted 1,273 ratings. The analysis uses the eight participants who completed all 150 assignments, yielding 1,200 ratings: 600 per model, with 200 per view count. The other six participants submitted 73 ratings in total; these partial responses are excluded from the statistics below. Each rating contains both an Action and a World score.

The completed assignments cover 198 distinct two-view source cases, 196 four-view cases, and 113 eight-view cases, identically paired across model variants. Thus, 600 ratings per model correspond to 507 distinct view-count/source-case combinations, with some cases rated by multiple participants. Table 8 reports pooled statistics for all five models, while Table 9 gives Khora and ablation category counts, means, and sample standard deviations by view count. Their pooled means weight the three view counts equally because each contributes 200 ratings.

Results and aggregation. Khora has Action 3.570000 and World 4.213333, which round to 3.570 and 4.213. Their arithmetic mean is 3.891667, reported as 3.892. The ablation has Action 2.075000 and World 4.065000, giving Overall 3.070000. Figure 9 visualizes the same 600 ratings per model. Participants and source cases contribute repeated observations, so the descriptive SDs and rating distributions do not establish statistical significance or an independent-sample confidence interval.

D Cross-View World Consistency

Our scene-level scores combine point-cloud geometry and color, as in the CVC formulation in MetaWorld's Appendix B.2 (Hu et al., 2026b). We refer specifically to that point-cloud definition, not a feature-similarity score. The two protocols below specify our input pairings, reconstruction, registration, and aggregation; neither is a reproduction of MetaWorld's evaluation protocol.

Table 9: **Khora human ratings by view count.** Eight participants who completed all assignments collectively contribute $n = 200$ ratings per model and view count. Columns 1–5 give exact counts; SD is descriptive. Bold marks the higher mean within each view count and criterion.

Method	Views	Criterion	Mean ↑	SD	1	2	3	4	5
Khora w/o Cross-Agent State	2	Action	2.095000	1.250517	87	54	25	21	13
Khora w/o Cross-Agent State	2	World	4.060000	1.015656	5	12	31	70	82
Khora w/o Cross-Agent State	4	Action	2.085000	1.263231	87	58	21	19	15
Khora w/o Cross-Agent State	4	World	4.085000	0.813069	0	9	31	94	66
Khora w/o Cross-Agent State	8	Action	2.045000	1.057624	79	59	38	22	2
Khora w/o Cross-Agent State	8	World	4.050000	0.860933	4	6	27	102	61
Khora	2	Action	**3.825000**	1.136147	11	15	38	70	66
Khora	2	World	**4.245000**	0.989886	5	9	22	60	104
Khora	4	Action	**3.530000**	1.036568	3	34	56	68	39
Khora	4	World	**4.185000**	0.880141	1	8	32	71	88
Khora	8	Action	**3.355000**	1.088531	7	44	50	69	30
Khora	8	World	**4.210000**	0.830269	1	6	28	80	85

D.1 CVC-GT: Measurement Protocol

CVC-GT is a GT-referenced scene-level proxy: generated and GT videos are reconstructed separately for each ego view, and the resulting point clouds are aligned and compared before averaging the two view scores. It does not directly register one generated view to another. Depth Anything V2 Small (Yang et al., 2024a) estimates depth; temporal ICP (Besl and McKay, 1992) accumulates each sequence, and point-cloud alignment precedes geometry and color comparison. The pipeline uses native output resolution, a 32-frame scoring window sampled every four frames (eight sampled frames per view), and image pixel stride eight. CVC, F-score, ColorSim, precision, and recall are averaged over valid cases. Baseline evaluations exclude videos rejected by their quality checks, so their means are conditional on the reported valid sets.

Solaris uses 100/100 valid/attempted cases, Gamma-World uses 9,988/9,988, and MIRA uses 100/102. Khora and its ablation use a paired 1,000/1,000-case evaluation with the same checkpoint and five denoising steps. Their sampled relative frames are 13, 17, 21, 25, 29, 33, 37, and 41. Baseline scoring windows start at saved-video frame 1 for Solaris and Gamma-World and 38 for MIRA. The action evaluation uses explicit source-frame mapping and separate comparison windows.

D.2 CVC-GT: Complete Results

Table 10 reports all components of the geometry-aware scores summarized in Table 3.

D.3 CVC-XView: Generated-View Comparison

CVC-XView compares generated view 0 with generated view 1 from the same case and model, without GT images, GT depth, or GT point clouds. It measures scene-level agreement between reconstructed generated observations, whereas CVC-GT measures agreement with a reference observation. Both retain the geometric and color components of the CVC implementation; the reconstruction, normalization, direction averaging, and cohorts differ, so their absolute scores must not be pooled or interpreted as a controlled protocol ablation.

Evaluation cohort. All five models are evaluated on the same 10,000 held-out cases, with two generated views per case. No cases are excluded based on generated image quality or scores, including

Table 10: **Complete CVC-GT results.** CVC and its components are GT-referenced geometric proxies with the model-specific protocols described above. Higher is better; six decimals are shown. ColorSim denotes color similarity.

Method	CVC-GT ↑	F-score ↑	ColorSim ↑	Precision ↑	Recall ↑
Solaris (Savva et al., 2026)	0.918310	0.940273	0.896348	0.956918	0.931426
Gamma-World (Liu et al., 2026)	0.897661	0.927937	0.867384	0.940621	0.923025
MIRA (Hu et al., 2026a)	0.922486	0.942951	0.902021	0.952500	0.940217
Khora w/o Cross-Agent State	<u>0.931900</u>	<u>0.959991</u>	<u>0.903809</u>	<u>0.963585</u>	<u>0.959753</u>
Khora	**0.932200**	**0.960586**	**0.903815**	**0.964424**	**0.960035**

low-motion and spatially flat outputs. All cases are successfully scored for every model, and results are averaged over this complete cohort rather than the GT-filtered 1,643-case cohort used for action evaluation.

Frame mapping and reconstruction. Each model uses a 32-frame generated window, sampled every four frames, giving eight RGB frames per view. All inputs are resized to 640×360. Physical source-frame indices are 2:4:30 for Solaris (Savva et al., 2026), 1:4:29 for Gamma-World (Liu et al., 2026), 38:4:66 for MIRA (Hu et al., 2026a), and 13:4:41 for Khora and its input ablation; $a : d : b$ denotes start, stride, and last sampled frame. The two views of a case use matching physical times, but the model families do not share a common physical scoring interval.

Depth Anything V2 Small (Yang et al., 2024a) predicts relative disparity q. With m the median positive disparity in that frame, we use $d = m/\max(q, 0.05m)$ and retain depths up to eight. Back-projection assumes a 90° horizontal field of view and samples image pixels with stride eight. Each view is reconstructed independently by accumulating its sampled frames using PCA-initialized, point-to-point ICP (Besl and McKay, 1992), with correspondence distance 0.4 and at most 20 iterations. Each accumulated cloud is centered at its bounding-box midpoint and divided by its largest extent. The two clouds are then aligned using PCA initialization followed by at most 20 iterations of point-to-point ICP, with correspondence distance 0.2. This differs from the FPFH-based global registration and point-to-plane ICP described in MetaWorld's Appendix B.2 (Hu et al., 2026b); our scoring also independently registers both directions, as specified below.

Components and aggregation. For each registration direction, let A be the aligned source cloud and B the reference cloud. Precision is the fraction of points in A whose nearest neighbor in B is within $\tau = 0.05$; recall reverses the point sets. Their harmonic mean is the F-score, set to zero when precision and recall sum to zero. ColorSim averages $1 - \|c_A - c_B\|_2/\sqrt{3}$ over source-to-reference nearest-neighbor pairs within τ, with RGB colors in $[0, 1]$; it is zero if no pair matches. For each direction, $\mathrm{CVC} = \frac{1}{2}(\text{F-score} + \mathrm{ColorSim})$. We independently register and score both directions, then average each component over directions and equally over the common cases. All five reported components lie in $[0, 1]$. F-score is computed before aggregation, so its reported mean need not equal the harmonic mean of mean precision and mean recall.

D.4 CVC-XView: Complete Results and Scope

Table 11 contains all five mean scores underlying the CVC-XView column of Table 3. These are descriptive means; differences do not by themselves establish statistical significance.

Interpretation and specificity. Independent scale normalization and unconstrained registration can align similar backgrounds while concealing incorrect pose, scale, or moving entities. There is no verified true-overlap mask or semantic identity matching. Consequently, this metric is a registration-based scene proxy; it does not establish synchronized actions or character identity. Table 12 reports a ten-case specificity control: replacing view 1 with the next case's view 1 can still produce high scores. These controls use a separate ten-case cohort and must not be mistaken for the 10,000-case model means. The GT-referenced action metrics in Appendix E provide complementary actor evidence.

Table 11: **Complete CVC-XView results.** All five models are evaluated on the same 10,000 cases. Higher is better; six decimals are shown. Physical time windows differ across model families.

Method	CVC-XView ↑	F-score ↑	ColorSim ↑	Precision ↑	Recall ↑
Solaris (Savva et al., 2026)	0.836662	0.841934	0.831390	0.904104	0.810266
Gamma-World (Liu et al., 2026)	0.836490	**0.870085**	0.802894	**0.919766**	**0.843761**
MIRA (Hu et al., 2026a)	0.830889	0.822776	**0.839002**	0.893388	0.788351
Khora w/o Cross-Agent State	0.838207	0.838615	0.837798	0.902333	0.806227
Khora	**0.838773**	0.839772	0.837774	0.903197	0.807770

Table 12: **Cross-case mismatch control for generated-view CVC.** Ten cases per model. Gamma-World (Liu et al., 2026) is compared with Khora. A high mismatched score exposes limited specificity of registration-based scene proxies.

Method	Pairing	n	CVC ↑	F-score ↑	ColorSim ↑	Precision ↑	Recall ↑
Gamma-World	paired	10	0.835207	0.871595	0.798820	0.927648	0.833106
Gamma-World	mismatch	10	0.797549	0.843013	0.752084	0.903676	0.802885
Khora	paired	10	0.850039	0.866512	0.833567	0.928074	0.838546
Khora	mismatch	10	0.832969	0.862446	0.803492	0.896035	0.843299

E Cross-View Action Consistency

E.1 Evaluation Protocol and Scope

E.1.1 Evaluation object and generation conditions

The quantitative action evaluation measures the other synchronized player's observable location, image motion, body configuration, and appearance in one generated ego view. Dataset identity and visibility associate the reference player; the evaluator does not compare two generated player tracks directly. This is a GT-referenced operationalization of cross-view action consistency. It does not recognize jump, crouch, fire, or reload events, and source action labels are never substituted for generated-video predictions.

Khora and the ablation use the same renderer checkpoint, seed, five denoising steps, initial 13 RGB frames, and video encoding. The ablation clears the non-ego actor geometry and state condition maps and pads the non-ego player-model-ID map at inference; scene conditions, ego conditions, and RGB history are retained. Thus the history can still show other players. The ablated model is not separately retrained. This evaluation uses recorded camera poses and supplied state trajectories to measure state-conditioned rendering, rather than autonomous prediction of future states.

E.1.2 Cohort construction

We apply the eligibility criteria to all 10,000 cases in the fixed two-player evaluation set, which is disjoint from the training set. One deterministic ego view is selected per case: the first view passing the metadata and geometric checks is tested for GT detectability, without falling back to the second view if that test fails. Both players must be alive over frames 13–57, the ego camera must be unscoped, the target must be visible at all 40 scored frames, and no third player may be visible in those frames. The target's projected body axis must remain in front of the camera, within the image margins, and have height 28–250 pixels at 640×360. Its ground-plane displacement between the endpoints of each window must be at least eight source-world units. A GT person detection associated with the projected target is required in every frame of both windows. Selection does not use generated images or scores.

Of the 10,000 cases, 7,541 fail the metadata/geometric checks, 816 fail the all-frame GT detection check, and 1,643 complete all available model/window evaluations, with no processing errors. This detector-observable, single-visible-target cohort is 16.43% of the source cases. It does not evaluate invisible-target hallucinations, multi-person identity switching, or all source trajectories without

Table 13: **Generation settings for action evaluation.** Diffusion steps describe the evaluated videos; model-family algorithms differ. Khora and its input ablation share all settings except other-agent conditions.

Method	RGB prefix	Diffusion steps	Native $W \times H$	Scored frames
Solaris (Savva et al., 2026)	1	50	640×352	13–32
Gamma-World (Liu et al., 2026)	1	35	480×320	Both windows
MIRA (Hu et al., 2026a)	38	50	512×288	38–57
Khora w/o Cross-Agent State	13	5	640×360	Both windows
Khora	13	5	640×360	Both windows

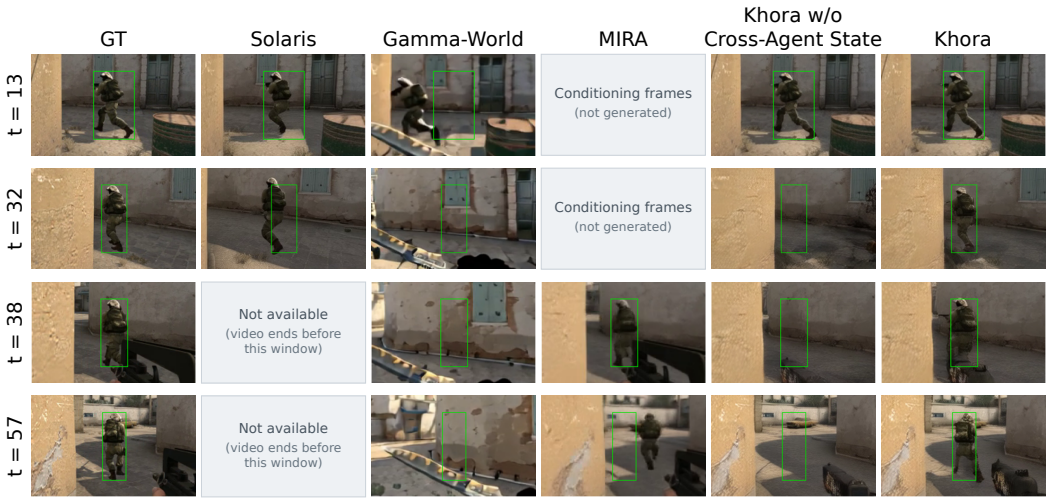

Figure 10: **Cross-view action consistency in fixed GT regions.** Case 79 is an example from the sampled inspection set. All methods use identical GT-defined crops and green reference boxes without generated recentering. Endpoint frames of each window are shown; numerical scores use all 20 frames at the original scoring resolution. Solaris is available only in the early window and MIRA only in the late window. All rows retain GT and five fixed model columns; gray cells identify conditioning frames or video coverage unavailable for generation scoring.

selection. The selected cases come from 1,466 source episodes. Every model/window contains 32,860 scored frames.

E.1.3 Time and spatial mapping

All indices are zero-based physical source-frame indices at 32 frames/s, not timestamps inferred from the encoded video's playback rate. A relative frame t maps to GT frame $b + t$, where b is the case's start frame. For Solaris, generated-frame index $t - 1$ corresponds to physical frame t; for the other models, generated-frame index t corresponds to physical frame t. We evaluate frames 13–32 for Solaris, Gamma-World, Khora, and its ablation, and frames 38–57 for MIRA, Gamma-World, Khora, and its ablation. Solaris ends before MIRA's 38-frame input prefix ends, so no common purely generated interval exists across all five models. Both windows contain 20 frames; they are short-window evaluations, not evidence of long-horizon stability. Equal physical frames within a panel do not equalize input history or state information across model families.

Scoring uses 640×360 images. Solaris is placed in rows 4–355 with four pixels of vertical padding on either side; other native outputs and GT are resized using area interpolation. GT scoring regions are clipped to the shared support $0 \leq x < 640$, $4 \leq y < 356$. No spatial registration, best-matching generated crop, or temporal warping is applied to action scores.

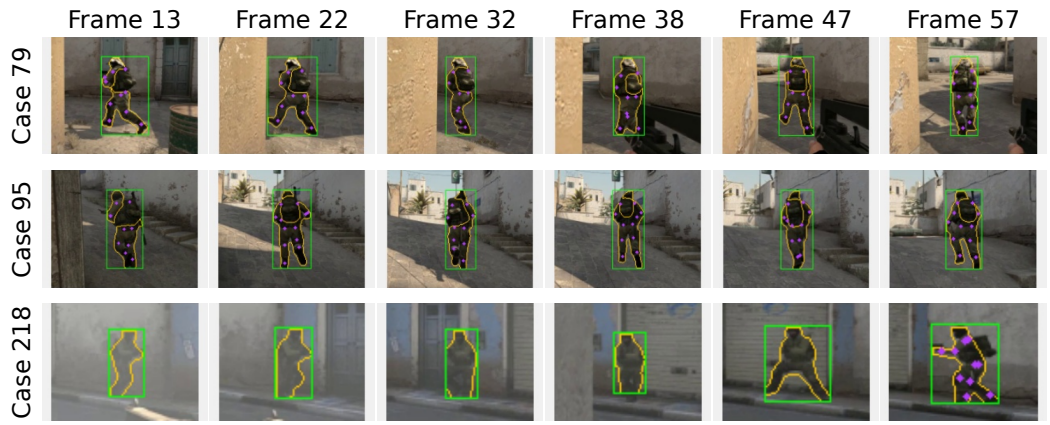

Figure 11: **GT box, foreground, and pose diagnostics.** Cases 79, 95, and 218 are sampled evaluation cases. Green outlines denote the fixed GT boxes; yellow contours denote foreground masks; magenta points denote body joints with confidence at least 0.5. The same three frames from each window are shown for every case. Crops are enlarged without altering the pseudo-labels.

E.1.4 GT PSEUDO-LABELS AND DETECTIONS

We use Ultralytics 8.3.203 (Jocher et al., 2023) with YOLOv8x detection, YOLOv8x-pose, and YOLOv8x-seg weights, inference size 960, FP32, and batch size four. Dataset player identity, world coordinates, rendering transform, and ego camera parameters provide a projected body axis to associate GT detections. The boxes, foreground masks, and 2D skeletons themselves are image-derived pseudo-labels, not native renderer masks or annotated skeletons. Detection confidence is 0.25 and NMS IoU is 0.6. The same fixed geometric filter is used for GT and generated detections: width at least five pixels, height at least 20 pixels, width/height at most 1.35, and exclusion of boxes with bottom at least 350 and top at least 216 to suppress first-person arms.

For the foreground, we run segmentation only on GT. A person mask must have confidence at least 0.25 and mask-box IoU at least 0.3 with the associated GT detection. Its intersection with the fixed GT box must contain at least 32 pixels and cover at least 10% of that box. Otherwise its foreground score is unavailable; the box-region score is still computed. Valid foreground coverage is 32,754/32,860 frames in the early window and 32,778/32,860 in the late window, identical for all methods within a window.

Figure 11 shows three sampled cases with the image-derived reference regions. The contours and keypoints illustrate the pseudo-label support used by the evaluator, rather than manually verified segmentation or pose.

E.1.5 DETECTION, POSITION, AND TRACKING

For the single GT target, a frame contributes one TP if the largest retained person-box IoU is at least 0.5. All other retained detections count as FP, and an unmatched GT target counts as FN. Precision, recall, and F1 pool these counts over cases. Mean best IoU includes missing frames with value zero. The separate single-target tracker initializes using the nearest eligible generated detection within $\max(60, h)$ pixels of the GT center, where h is GT box height. Initialization can occur after the first frame; after the first association it uses only the last generated box, with no GT updates or interpolation of missing frames.

Let $D = \sqrt{640^2 + 360^2}$, and let e_t be the center error divided by D. Center PCK counts $e_t \leq 0.05$, with missing tracks incorrect. Position AUC is the mean of $\max(0, 1 - e_t/0.1)$, with missing tracks assigned zero. ADE averages errors over available tracked frames in each case, and FDE uses the last frame only when a track is available there. The word "available" does not imply a correct IoU match. Their missing-penalized counterparts assign missing frames one image diagonal and cap available errors at that value. Normalized errors are reported as percentages of D. Track coverage, track breaks, and excess detections per frame accompany the conditional errors to expose survivorship effects.

E.1.6 IMAGE MOTION AND BODY POSE

Motion pairs are four frames apart. A frame pair is classified as moving if the GT center displacement is at least three pixels. Direction is correct when the generated displacement is also at least three pixels and its cosine with GT is at least 0.5; missing tracks are incorrect. Direction accuracy pools correct counts over GT-moving pairs, while image move/stop accuracy pools all pairs. The available-track cosine is a separate conditional statistic. Motion includes ego-camera motion and is not world-space trajectory accuracy. Motion-pair coverage is the fraction of frame pairs with an available generated track at both endpoints. Higher coverage indicates greater track availability, not necessarily correct motion, and is read alongside direction and move/stop accuracy. Tracking and generated-joint coverage are ranked as model-dependent statistics; fixed GT box and foreground coverage describe the evaluation regions and are not ranked.

Pose scoring uses COCO (Lin et al., 2014) joints 5–16 (shoulders through ankles), retaining GT joints with confidence at least 0.5. A generated joint must also have confidence at least 0.5 and lie within $0.2h$ of the GT joint for absolute PCK; unavailable predictions count as incorrect. PCK pools correct and eligible GT joints. Root-normalized shape PCK separately subtracts each pose's hip center and divides by its own pose-box height, with threshold 0.2; it requires confident GT hips and treats unavailable predicted hips as incorrect. This supplementary score removes location and scale information.

E.1.7 FIXED-REGION PIXEL METRICS

For a GT box or valid GT foreground R_t, we compute

$$\mathrm{MAE}_t = \frac{1}{3|R_t|} \sum_{p \in R_t} \sum_{c=1}^{3} \frac{|I^{\mathrm{gen}}_{t,p,c} - I^{\mathrm{GT}}_{t,p,c}|}{255}, \qquad S_t = 1 - \mathrm{MAE}_t. \tag{19}$$

Table 4 reports foreground MAE (lower is better). The complete tables and diagnostic controls also report the equivalent similarity $S_t = 1 - \mathrm{MAE}_t$. Its bootstrap interval $[L, U]$ becomes $[1 - U, 1 - L]$ for MAE; paired Khora-minus-ablation differences change sign, with interval $[-U, -L]$. The same GT-defined region is used for every generated model, including frames with no detected generated person. MSE uses squared normalized RGB differences, and PSNR is $\min(100, -10 \log_{10}(\max(\mathrm{MSE}, 10^{-10})))$. Exact-match frame counts are retained. SSIM (Wang et al., 2004) uses an 11×11 Gaussian window with $\sigma = 1.5$, population variance, and constants 0.01^2 and 0.03^2. RGB-channel SSIM maps are averaged and pooled only inside the GT region; the local SSIM window includes up to five context pixels outside it. SSIM retains its raw range, including negative values. The color score $1 - \mathrm{MAE}$ is sensitive to brightness and can be high even when the person is absent.

E.1.8 AGGREGATION AND UNCERTAINTY

Scalar errors and pixel scores first average valid frames within a case and then average valid cases equally. A PSNR aggregate therefore averages frame PSNR values; it is not PSNR of the aggregate MSE. Detection P/R/F1, direction, move/stop accuracy, pose PCK, and joint coverage pool their appropriate counts. Counts are summed, except track breaks, which are a per-case mean. The complete tables also report available-case denominators, across-case sample standard deviations (with $n - 1$ denominator), and supplementary frame-weighted pixel similarity and SSIM. A case SD describes the distribution of per-case ratios, even when the primary statistic pools counts; it is not a standard error of that pooled statistic.

We compute 2,000 paired bootstrap (Efron, 1979) replicates (seed 20260925), sampling the 1,466 source episodes with replacement and retaining all selected cases from each sampled episode. Each episode contributes at most four cases. The same resamples are used for all models and windows. Ratio statistics are recomputed from resampled counts. Tables 17 and 18 give percentile 95% intervals for the main scores and Khora-minus-ablation differences. These intervals quantify sampling uncertainty conditional on this cohort; they do not account for pseudo-label bias or model-training randomness.

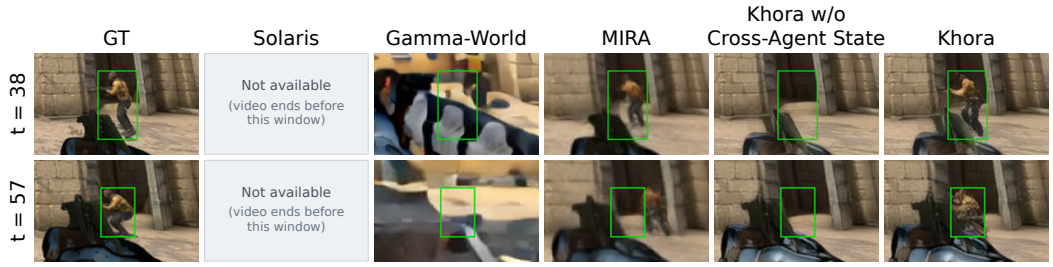

Figure 12: **A visible actor missed by the detector.** Case 3629 is selected because it has the lowest late-window Khora F1 score in the random inspection set. Despite zero detector F1, a player remains visible: the frame-38 confidence is 0.1445, below the 0.25 threshold. All methods use the same enlarged GT crop and green reference box. This diagnoses detector sensitivity, not verified actor absence. The six columns match Figure 10; the Solaris column is marked unavailable for this window.

E.1.9 Metric controls and limitations

Identity, missing-person, spatial-shift, time-reversal, and frozen-frame controls check metric behavior. A black-image control can still yield foreground $1 - \mathrm{MAE} = 0.864479$ while foreground SSIM is 0.013803, illustrating why pixel similarity cannot establish actor presence. The control is illustrative, not an aggregate baseline. We also visually inspect 600 GT crops (three frames in each window for 100 cases sampled uniformly without replacement, seed 20260925), including box, mask, and confident-joint overlays. A 20-case subset additionally supplies 160 generated midpoint crops across all available model/window combinations. Small targets often lack confident body joints, and segmentation boundaries can include equipment or omit fine limbs. These visual diagnostics are not an independent human annotation study or an estimate of annotator agreement. Detector and pose-estimator domain bias remains, especially for small, blurred, or stylized generated people. Semantic jump/crouch/fire/reload accuracy, multi-target identity metrics, and world-space motion accuracy are not computed and have no numerical entries.

The low-F1 example in Figure 12 also exposes detector false negatives on a visible generated player. To examine threshold sensitivity, we recompute detection P/R/F1 on all 1,643 cases at confidence thresholds 0.15, 0.25, 0.35, and 0.50, holding GT references and geometric filters fixed. The within-window F1 ordering is unchanged at all four thresholds. Khora F1 ranges from 0.888349 to 0.904063 in the early window and 0.880531 to 0.902400 in the late window. This check does not eliminate detector domain bias. Table 19 reports every count and P/R/F1 value from this sensitivity analysis.

E.2 Diagnostics and Visual Summaries

Figure 13 visualizes three complementary quantities from the complete confidence-interval table: person localization, image-motion direction, and appearance in the fixed GT foreground. Each window uses the same 1,643 selected cases. The two windows remain separate, and absent model/window combinations are not imputed. Confidence intervals are computed by resampling source episodes, preserving within-episode dependence; they address sampling uncertainty rather than detector bias.

E.2.1 Detector sensitivity

Figure 14 varies only the confidence threshold applied to generated detections. GT pseudo-labels, case selection, and geometric filtering remain fixed. All four thresholds preserve the within-window F1 ordering, but threshold stability does not validate every individual detection. In particular, the visible player in the low-F1 example can be missed by the detector. The complete threshold table reports all TP, FP, FN, precision, recall, and F1 entries.

E.2.2 Pixel-score controls

We use one diagnostic clip for identity, black, white, time reversal, frozen-first-frame, 100-pixel shift, and GT-actor-inpainting controls. Table 14 reports appearance scores for both the GT box and GT

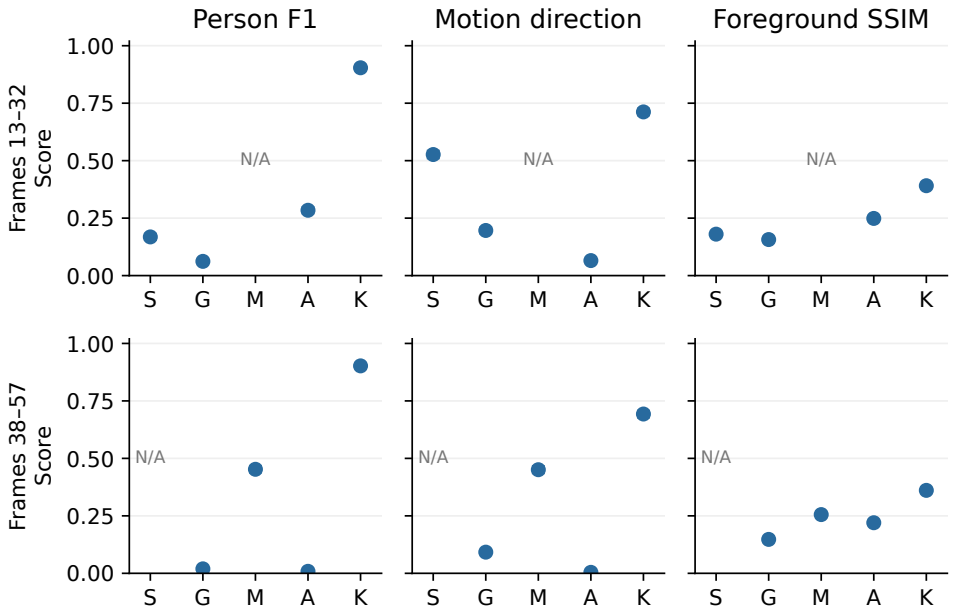

Figure 13: **Action consistency with episode-cluster uncertainty.** S: Solaris; G: Gamma-World; M: MIRA; A: Khora w/o Cross-Agent State; K: Khora. Fixed model positions follow publication order, with N/A marking unavailable model/window combinations. Dots show estimates; bars show 95% percentile intervals from 2,000 paired episode bootstrap replicates. Higher values are better for all displayed metrics. Exact values, the other main metrics, and paired differences appear in Tables 17 and 18.

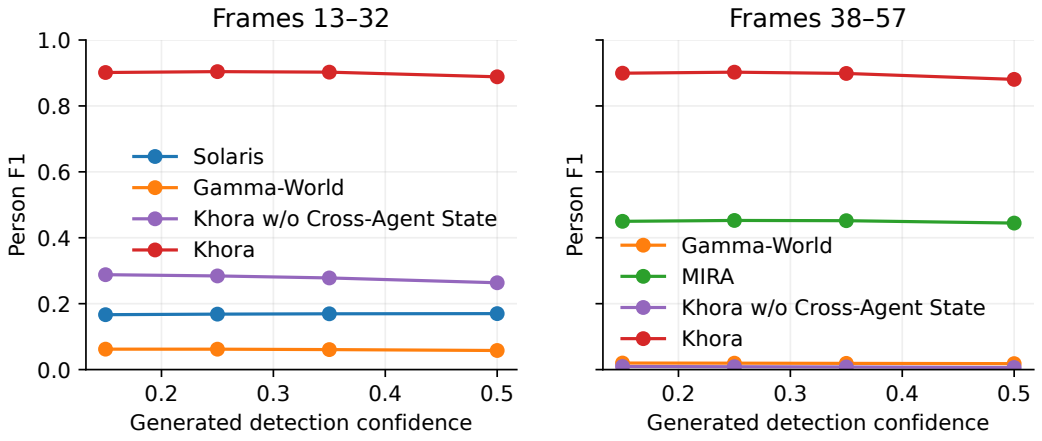

Figure 14: **Sensitivity of person F1 to detection confidence.** The two panels use different physical frame windows. Lines join the four recomputed thresholds, 0.15, 0.25, 0.35, and 0.50; the main evaluation uses 0.25. These curves reuse all 1,643 cases and do not reselect the cohort.

Table 14: **All fixed-region appearance controls.** One diagnostic clip, 20 frames; 313,284 box pixels and 126,508 foreground pixels. All regions are valid in all frames. Identity has 20 exact matches; frozen-first-frame has one; other controls have zero. PSNR is capped at 100 dB.

Control	Region	MAE ↓	MSE ↓	PSNR ↑	1−MAE ↑	SSIM ↑
identity	box	0.000000	0.000000	100.000000	1.000000	1.000000
identity	foreground	0.000000	0.000000	100.000000	1.000000	1.000000
black	box	0.193999	0.052088	12.939865	0.806001	0.009516
black	foreground	0.135521	0.024501	16.161390	0.864479	0.013803
white	box	0.806001	0.664090	1.779099	0.193999	0.184863
white	foreground	0.864479	0.753459	1.231538	0.135521	0.113299
reverse time	box	0.121000	0.027501	16.333985	0.879000	0.310654
reverse time	foreground	0.132385	0.027853	16.037519	0.867615	0.253468
frozen first frame	box	0.094805	0.018044	22.386295	0.905195	0.366631
frozen first frame	foreground	0.101878	0.017520	22.081798	0.898122	0.293297
shift100px	box	0.153501	0.043562	14.393723	0.846499	0.264591
shift100px	foreground	0.159509	0.045927	14.203725	0.840491	0.225933
Actor inpainted	box	0.042571	0.006973	21.947406	0.957429	0.713602
Actor inpainted	foreground	0.103889	0.017032	18.020472	0.896111	0.383361

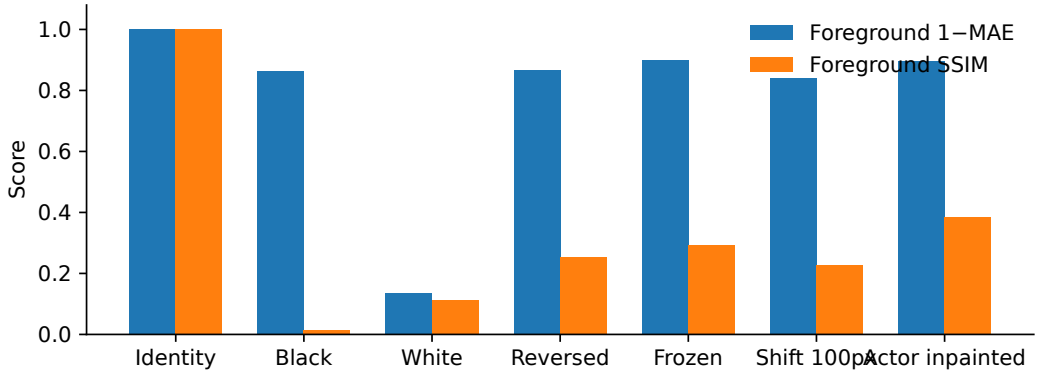

Figure 15: **Why fixed-region RGB similarity needs structural checks.** All bars use the same GT foreground. A black prediction or an inpainted actor can retain a high $1 - \mathrm{MAE}$ score. SSIM is more sensitive to local structure but also cannot certify actor presence by itself. The identity control reaches one on both scores.

foreground. All control regions are fixed from GT; no generated region is selected or aligned. These controls test metric behavior on one clip and are not extra model baselines or population means.

The controls motivate joint reporting rather than treating color similarity as a detection accuracy. Pixel metrics deliberately retain frames where the generated person is absent; localization, pose, and motion metrics expose errors that a background-dominated or brightness-sensitive average may hide.

PSNR averages the per-frame logarithmic scores, whereas MSE is averaged in linear RGB-error space. Consequently, the printed mean PSNR need not equal $-10 \log_{10}$ of the printed mean MSE. Exact-match frames use the same 100 dB cap as the full model evaluation.

E.3 Complete Numerical Results

Tables 15 and 16 report the complete results across eight model/window combinations, including supplementary scores, valid-case counts, and sample standard deviations. Main-text values use four decimals; these full tables use six. All methods share the same selected cases within each window. "N/A" denotes an undefined statistic, never a zero imputed for a failed GT pseudo-label.

Table 15: **Complete cross-view action consistency results, frames 13–32.** All scalar metrics, counts, valid-case denominators, and sample standard deviations. Six decimal places are shown; integer counts are exact. Baselines: Solaris (Savva et al., 2026) and Gamma-World (Liu et al., 2026).

Statistic	Solaris	Gamma-World	Khora w/o Cross-Agent State	Khora
Detection, position, and trajectory				
Available-track ADE ($\%D$) ↓	5.155349	7.332127	1.510989	**0.532387**
Available-track ADE (pixels) ↓	37.855851	53.839983	11.095231	**3.909328**
Missing-penalized ADE ($\%D$) ↓	20.379866	53.844661	74.050573	**13.032861**
Direction accuracy (4 frames) ↑	0.526670	0.196322	0.065353	**0.712070**
Available-track direction cosine ↑	0.515515	0.278088	0.313948	**0.830552**
Extra detections per frame ↓	0.041570	0.085027	**0.012203**	0.039623
Detection F1 ↑	0.168277	0.061932	0.284192	**0.904063**
Available-track FDE (pixels) ↓	52.059019	67.112213	42.719948	**3.987530**
Missing-penalized FDE ($\%D$) ↓	23.941264	58.745059	91.860086	**14.526335**
Mean best IoU, all frames ↑	0.204078	0.084616	0.158711	**0.702175**
Motion-pair coverage ↑	0.801430	0.398509	0.165056	**0.809457**
Image move/stop accuracy ↑	0.645123	0.311967	0.115110	**0.722573**
Center PCK ($0.05D$) ↑	0.532988	0.215307	0.241996	**0.871029**
Position AUC (0–$0.1D$) ↑	0.463675	0.204899	0.216242	**0.828912**
Detection precision ↑	0.176077	0.075568	0.612741	**0.938466**
Detection recall ↑	0.161138	0.052465	0.184997	**0.872094**
Target-track frame coverage ↑	0.839166	0.496287	0.264699	**0.874254**
Track breaks per case ↓	**0.468655**	1.294583	0.914790	0.768107
Pose scores				
Absolute pose PCK ($0.2h$) ↑	0.211109	0.049990	0.172258	**0.743765**
Generated joint coverage ↑	0.719683	0.360884	0.225433	**0.772633**
Root-normalized shape PCK ↑	0.679222	0.326775	0.211646	**0.744580**
Fixed GT box and foreground appearance				
Box frame coverage	1.000000	1.000000	1.000000	1.000000
Box MAE ↓	0.137810	0.174213	0.100830	**0.069708**
Box $1 - \mathrm{MAE}$ ↑	0.862190	0.825787	0.899170	**0.930292**
Box MSE ↓	0.040059	0.059326	0.026387	**0.013350**
Box PSNR (dB, cap 100) ↑	15.899162	14.136474	17.852195	**20.411451**
Box SSIM ↑	0.248882	0.212112	0.388850	**0.473400**
Box $1 - \mathrm{MAE}$, frame-weighted ↑	0.862190	0.825787	0.899170	**0.930292**
Box SSIM, frame-weighted ↑	0.248882	0.212112	0.388850	**0.473400**
Foreground frame coverage	0.996774	0.996774	0.996774	0.996774
Foreground MAE ↓	0.164680	0.199953	0.141726	**0.083418**
Foreground $1 - \mathrm{MAE}$ ↑	0.835320	0.800047	0.858274	**0.916582**
Foreground MSE ↓	0.050716	0.072776	0.040470	**0.016864**
Foreground PSNR (dB, cap 100) ↑	15.055761	13.481165	16.134487	**19.359696**
Foreground SSIM ↑	0.180287	0.156796	0.248728	**0.391012**
Foreground $1 - \mathrm{MAE}$, frame-weighted ↑	0.835348	0.800036	0.858299	**0.916589**
Foreground SSIM, frame-weighted ↑	0.180345	0.156823	0.248720	**0.390966**

Table 15: **Continued.** Complete cross-view action consistency results, frames 13–32.

Statistic	Solaris	Gamma-World	Khora w/o Cross-Agent State	Khora
Counts and coverage denominators				
Cases	1643	1643	1643	1643
False negatives (IoU ≥ 0.5)	27565	31136	26781	4203
False positives (IoU ≥ 0.5)	24777	21090	3842	1879
Correct-direction pairs	12056	4494	1496	16300
Correct move/stop pairs	16959	8201	3026	18995
Scored frames	32860	32860	32860	32860
GT moving pairs	22891	22891	22891	22891
All motion pairs	26288	26288	26288	26288
Retained generated detections	30072	22814	9921	30536
True positives (IoU ≥ 0.5)	5295	1724	6079	28657
Correct absolute body joints	51643	12229	42139	181945
Confident GT body joints	244627	244627	244627	244627
Frames with GT pose	23028	23028	23028	23028
Available generated body joints	176054	88282	55147	189007
Frames with paired generated pose	17897	9383	5664	18849
Correct normalized shape joints	166152	79936	51773	182140
GT joints eligible for shape	244621	244621	244621	244621
Box valid frames	32860	32860	32860	32860
Box candidate frames	32860	32860	32860	32860
Box scored pixels	139719152	139719152	139719152	139719152
Box exact-match frames	0	0	0	0
Foreground valid frames	32754	32754	32754	32754
Foreground candidate frames	32860	32860	32860	32860
Foreground scored pixels	62380972	62380972	62380972	62380972
Foreground exact-match frames	0	0	0	0
Valid-case counts for each statistic				
Available-track ADE ($\%D$): valid cases	1525	1259	1241	1634
Available-track ADE (pixels): valid cases	1525	1259	1241	1634
Missing-penalized ADE ($\%D$): valid cases	1643	1643	1643	1643
Direction accuracy (4 frames): valid cases	1638	1638	1638	1638
Available-track direction cosine: valid cases	1477	1063	596	1587
Extra detections per frame: valid cases	1643	1643	1643	1643
Detection F1: valid cases	1643	1643	1643	1643
Available-track FDE (pixels): valid cases	1345	746	142	1412
Missing-penalized FDE ($\%D$): valid cases	1643	1643	1643	1643
False negatives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
False positives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
Mean best IoU, all frames: valid cases	1643	1643	1643	1643
Motion-pair coverage: valid cases	1643	1643	1643	1643
Image move/stop accuracy: valid cases	1643	1643	1643	1643
Correct-direction pairs: valid cases	1643	1643	1643	1643
Correct move/stop pairs: valid cases	1643	1643	1643	1643
Scored frames: valid cases	1643	1643	1643	1643
GT moving pairs: valid cases	1643	1643	1643	1643

Table 15: **Continued.** Complete cross-view action consistency results, frames 13–32.

Statistic	Solaris	Gamma-World	Khora w/o Cross-Agent State	Khora
Valid-case counts for each statistic (continued)				
All motion pairs: valid cases	1643	1643	1643	1643
Retained generated detections: valid cases	1643	1643	1643	1643
Center PCK ($0.05D$): valid cases	1643	1643	1643	1643
Position AUC (0–$0.1D$): valid cases	1643	1643	1643	1643
Detection precision: valid cases	1643	1643	1643	1643
Detection recall: valid cases	1643	1643	1643	1643
Target-track frame coverage: valid cases	1643	1643	1643	1643
True positives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
Track breaks per case: valid cases	1643	1643	1643	1643
Absolute pose PCK ($0.2h$): valid cases	1288	1288	1288	1288
Correct absolute body joints: valid cases	1643	1643	1643	1643
Confident GT body joints: valid cases	1643	1643	1643	1643
Frames with GT pose: valid cases	1643	1643	1643	1643
Generated joint coverage: valid cases	1288	1288	1288	1288
Available generated body joints: valid cases	1643	1643	1643	1643
Frames with paired generated pose: valid cases	1643	1643	1643	1643
Root-normalized shape PCK: valid cases	1288	1288	1288	1288
Correct normalized shape joints: valid cases	1643	1643	1643	1643
GT joints eligible for shape: valid cases	1643	1643	1643	1643
Box MAE: valid cases	1643	1643	1643	1643
Box $1 - \mathrm{MAE}$: valid cases	1643	1643	1643	1643
Box MSE: valid cases	1643	1643	1643	1643
Box PSNR (dB, cap 100): valid cases	1643	1643	1643	1643
Box SSIM: valid cases	1643	1643	1643	1643
Foreground MAE: valid cases	1643	1643	1643	1643
Foreground $1 - \mathrm{MAE}$: valid cases	1643	1643	1643	1643
Foreground MSE: valid cases	1643	1643	1643	1643
Foreground PSNR (dB, cap 100): valid cases	1643	1643	1643	1643
Foreground SSIM: valid cases	1643	1643	1643	1643
Across-case sample standard deviations				
Available-track ADE (pixels): case SD	31.217285	35.289713	14.888488	4.970599
Mean best IoU, all frames: case SD	0.209103	0.133646	0.172150	0.174980
Position AUC (0–$0.1D$): case SD	0.298586	0.242170	0.235268	0.205931
Absolute pose PCK ($0.2h$): case SD	0.275559	0.123267	0.202844	0.302854
Box MAE: case SD	0.069088	0.081146	0.045313	0.027811
Box $1 - \mathrm{MAE}$: case SD	0.069088	0.081146	0.045313	0.027811
Box MSE: case SD	0.034486	0.046501	0.021108	0.009809
Box PSNR (dB, cap 100): case SD	4.341592	4.299445	4.276162	3.827186
Box SSIM: case SD	0.135028	0.125549	0.111220	0.095963
Foreground MAE: case SD	0.086697	0.099556	0.068595	0.032630
Foreground $1 - \mathrm{MAE}$: case SD	0.086697	0.099556	0.068595	0.032630
Foreground MSE: case SD	0.047088	0.060868	0.034412	0.012490
Foreground PSNR (dB, cap 100): case SD	4.430126	4.455401	4.326201	3.726357
Foreground SSIM: case SD	0.112952	0.102911	0.111444	0.098158

Table 16: **Complete cross-view action consistency results, frames 38–57.** All scalar metrics, counts, valid-case denominators, and sample standard deviations. Six decimal places are shown; integer counts are exact. Baselines: Gamma-World (Liu et al., 2026) and MIRA (Hu et al., 2026a).

Statistic	Gamma-World	MIRA	Khora w/o Cross-Agent State	Khora
Detection, position, and trajectory				
Available-track ADE ($\%D$) ↓	8.480216	3.085277	6.758419	**0.572149**
Available-track ADE (pixels) ↓	62.270429	22.655263	49.627232	**4.201300**
Missing-penalized ADE ($\%D$) ↓	73.272929	17.046396	96.600790	**13.700354**
Direction accuracy (4 frames) ↑	0.092087	0.450706	0.004259	**0.692939**
Available-track direction cosine ↑	0.195565	0.361216	0.158781	**0.806403**
Extra detections per frame ↓	0.067985	0.033262	**0.002161**	0.035180
Detection F1 ↑	0.019482	0.452510	0.008690	**0.902400**
Available-track FDE (pixels) ↓	66.919357	41.816554	55.790681	**4.219938**
Missing-penalized FDE ($\%D$) ↓	76.987914	23.430781	97.637923	**19.515723**
Mean best IoU, all frames ↑	0.035263	0.391557	0.007622	**0.685430**
Motion-pair coverage ↑	0.221432	0.799871	0.025829	**0.804664**
Image move/stop accuracy ↑	0.165969	0.624011	0.014227	**0.703971**
Center PCK ($0.05D$) ↑	0.091023	0.702648	0.019781	**0.865520**
Position AUC (0–$0.1D$) ↑	0.096553	0.623204	0.017384	**0.819846**
Detection precision ↑	0.027567	0.477649	0.065023	**0.943756**
Detection recall ↑	0.015064	0.429884	0.004656	**0.864516**
Target-track frame coverage ↑	0.290809	0.856086	0.036062	**0.867833**
Track breaks per case ↓	0.926354	0.763238	**0.121120**	0.810712
Pose scores				
Absolute pose PCK ($0.2h$) ↑	0.018057	0.412928	0.004485	**0.738481**
Generated joint coverage ↑	0.200740	0.692774	0.021243	**0.767578**
Root-normalized shape PCK ↑	0.177811	0.652417	0.017866	**0.738853**
Fixed GT box and foreground appearance				
Box frame coverage	1.000000	1.000000	1.000000	1.000000
Box MAE ↓	0.201603	0.108183	0.109529	**0.073891**
Box $1 - \mathrm{MAE}$ ↑	0.798397	0.891817	0.890471	**0.926109**
Box MSE ↓	0.075867	0.026792	0.030498	**0.014456**
Box PSNR (dB, cap 100) ↑	13.167113	17.715944	17.079508	**19.967869**
Box SSIM ↑	0.195440	0.329751	0.381442	**0.445497**
Box $1 - \mathrm{MAE}$, frame-weighted ↑	0.798397	0.891817	0.890471	**0.926109**
Box SSIM, frame-weighted ↑	0.195440	0.329751	0.381442	**0.445497**
Foreground frame coverage	0.997505	0.997505	0.997505	0.997505
Foreground MAE ↓	0.226105	0.126754	0.162544	**0.086796**
Foreground $1 - \mathrm{MAE}$ ↑	0.773895	0.873246	0.837456	**0.913204**
Foreground MSE ↓	0.091118	0.032994	0.049062	**0.017780**
Foreground PSNR (dB, cap 100) ↑	12.664245	16.916130	15.030659	**19.040465**
Foreground SSIM ↑	0.147564	0.255510	0.220065	**0.360863**
Foreground $1 - \mathrm{MAE}$, frame-weighted ↑	0.773969	0.873271	0.837491	**0.913202**
Foreground SSIM, frame-weighted ↑	0.147587	0.255570	0.220087	**0.360813**

Table 16: **Continued.** Complete cross-view action consistency results, frames 38–57.

Statistic	Gamma-World	MIRA	Khora w/o Cross-Agent State	Khora
Counts and coverage denominators				
Cases	1643	1643	1643	1643
False negatives (IoU ≥ 0.5)	32365	18734	32707	4452
False positives (IoU ≥ 0.5)	17461	15448	2200	1693
Correct-direction pairs	2054	10053	95	15456
Correct move/stop pairs	4363	16404	374	18506
Scored frames	32860	32860	32860	32860
GT moving pairs	22305	22305	22305	22305
All motion pairs	26288	26288	26288	26288
Retained generated detections	17956	29574	2353	30101
True positives (IoU ≥ 0.5)	495	14126	153	28408
Correct absolute body joints	4421	101101	1098	180809
Confident GT body joints	244839	244839	244839	244839
Frames with GT pose	23005	23005	23005	23005
Available generated body joints	49149	169618	5201	187933
Frames with paired generated pose	5343	17099	651	18763
Correct normalized shape joints	43533	159730	4374	180892
GT joints eligible for shape	244828	244828	244828	244828
Box valid frames	32860	32860	32860	32860
Box candidate frames	32860	32860	32860	32860
Box scored pixels	141430664	141430664	141430664	141430664
Box exact-match frames	0	0	0	0
Foreground valid frames	32778	32778	32778	32778
Foreground candidate frames	32860	32860	32860	32860
Foreground scored pixels	63335210	63335210	63335210	63335210
Foreground exact-match frames	0	0	0	0
Valid-case counts for each statistic				
Available-track ADE ($\%D$): valid cases	847	1617	145	1631
Available-track ADE (pixels): valid cases	847	1617	145	1631
Missing-penalized ADE ($\%D$): valid cases	1643	1643	1643	1643
Direction accuracy (4 frames): valid cases	1624	1624	1624	1624
Available-track direction cosine: valid cases	650	1537	75	1562
Extra detections per frame: valid cases	1643	1643	1643	1643
Detection F1: valid cases	1643	1643	1643	1643
Available-track FDE (pixels): valid cases	416	1334	42	1330
Missing-penalized FDE ($\%D$): valid cases	1643	1643	1643	1643
False negatives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
False positives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
Mean best IoU, all frames: valid cases	1643	1643	1643	1643
Motion-pair coverage: valid cases	1643	1643	1643	1643
Image move/stop accuracy: valid cases	1643	1643	1643	1643
Correct-direction pairs: valid cases	1643	1643	1643	1643
Correct move/stop pairs: valid cases	1643	1643	1643	1643
Scored frames: valid cases	1643	1643	1643	1643
GT moving pairs: valid cases	1643	1643	1643	1643

Table 16: **Continued.** Complete cross-view action consistency results, frames 38–57.

Statistic	Gamma-World	MIRA	Khora w/o Cross-Agent State	Khora
Valid-case counts for each statistic (continued)				
All motion pairs: valid cases	1643	1643	1643	1643
Retained generated detections: valid cases	1643	1643	1643	1643
Center PCK ($0.05D$): valid cases	1643	1643	1643	1643
Position AUC (0–$0.1D$): valid cases	1643	1643	1643	1643
Detection precision: valid cases	1643	1643	1643	1643
Detection recall: valid cases	1643	1643	1643	1643
Target-track frame coverage: valid cases	1643	1643	1643	1643
True positives (IoU ≥ 0.5): valid cases	1643	1643	1643	1643
Track breaks per case: valid cases	1643	1643	1643	1643
Absolute pose PCK ($0.2h$): valid cases	1311	1311	1311	1311
Correct absolute body joints: valid cases	1643	1643	1643	1643
Confident GT body joints: valid cases	1643	1643	1643	1643
Frames with GT pose: valid cases	1643	1643	1643	1643
Generated joint coverage: valid cases	1311	1311	1311	1311
Available generated body joints: valid cases	1643	1643	1643	1643
Frames with paired generated pose: valid cases	1643	1643	1643	1643
Root-normalized shape PCK: valid cases	1311	1311	1311	1311
Correct normalized shape joints: valid cases	1643	1643	1643	1643
GT joints eligible for shape: valid cases	1643	1643	1643	1643
Box MAE: valid cases	1643	1643	1643	1643
Box $1 - \mathrm{MAE}$: valid cases	1643	1643	1643	1643
Box MSE: valid cases	1643	1643	1643	1643
Box PSNR (dB, cap 100): valid cases	1643	1643	1643	1643
Box SSIM: valid cases	1643	1643	1643	1643
Foreground MAE: valid cases	1643	1643	1643	1643
Foreground $1 - \mathrm{MAE}$: valid cases	1643	1643	1643	1643
Foreground MSE: valid cases	1643	1643	1643	1643
Foreground PSNR (dB, cap 100): valid cases	1643	1643	1643	1643
Foreground SSIM: valid cases	1643	1643	1643	1643
Across-case sample standard deviations				
Available-track ADE (pixels): case SD	36.328662	23.289835	36.659425	4.773047
Mean best IoU, all frames: case SD	0.082983	0.215464	0.043863	0.181199
Position AUC (0–$0.1D$): case SD	0.176210	0.250192	0.090624	0.215206
Absolute pose PCK ($0.2h$): case SD	0.070866	0.313627	0.036918	0.308651
Box MAE: case SD	0.098199	0.048697	0.051581	0.030293
Box $1 - \mathrm{MAE}$: case SD	0.098199	0.048697	0.051581	0.030293
Box MSE: case SD	0.061671	0.020717	0.025337	0.011033
Box PSNR (dB, cap 100): case SD	4.490447	4.081494	4.383445	3.776860
Box SSIM: case SD	0.121931	0.129327	0.113693	0.097813
Foreground MAE: case SD	0.118242	0.059475	0.078319	0.034689
Foreground $1 - \mathrm{MAE}$: case SD	0.118242	0.059475	0.078319	0.034689
Foreground MSE: case SD	0.078751	0.027252	0.040515	0.013399
Foreground PSNR (dB, cap 100): case SD	4.717251	4.057525	4.398476	3.688559
Foreground SSIM: case SD	0.101376	0.114871	0.108171	0.095651

Table 17: **Episode-cluster bootstrap intervals for all main-table scores.** Two thousand paired replicates over 1,466 source episodes; percentile 95% intervals, conditional on the selected evaluation cohort. Baselines: Solaris (Savva et al., 2026), Gamma-World (Liu et al., 2026), and MIRA (Hu et al., 2026a).

Method	Metric	Estimate	Lower 95%	Upper 95%
Relative frames 13–32				
Solaris	Detection F1 ↑	0.168277	0.155148	0.181531
	Detection recall ↑	0.161138	0.148115	0.174070
	Direction accuracy (4 frames) ↑	0.526670	0.510304	0.543214
	Absolute pose PCK ($0.2h$) ↑	0.211109	0.195324	0.227955
	Foreground SSIM ↑	0.180287	0.175056	0.186040
	Foreground MAE ↓	0.164680	0.160421	0.168658
Gamma-World	Detection F1 ↑	0.061932	0.053986	0.070030
	Detection recall ↑	0.052465	0.045684	0.059479
	Direction accuracy (4 frames) ↑	0.196322	0.183887	0.208817
	Absolute pose PCK ($0.2h$) ↑	0.049990	0.042895	0.057416
	Foreground SSIM ↑	0.156796	0.151948	0.161934
	Foreground MAE ↓	0.199953	0.194890	0.204760
Khora w/o Cross-Agent State	Detection F1 ↑	0.284192	0.271254	0.297233
	Detection recall ↑	0.184997	0.175137	0.194768
	Direction accuracy (4 frames) ↑	0.065353	0.058684	0.072398
	Absolute pose PCK ($0.2h$) ↑	0.172258	0.160885	0.183308
	Foreground SSIM ↑	0.248728	0.243653	0.254287
	Foreground MAE ↓	0.141726	0.138325	0.145015
Khora	Detection F1 ↑	0.904063	0.897255	0.910779
	Detection recall ↑	0.872094	0.862477	0.881669
	Direction accuracy (4 frames) ↑	0.712070	0.697602	0.725972
	Absolute pose PCK ($0.2h$) ↑	0.743765	0.730095	0.756911
	Foreground SSIM ↑	0.391012	0.386251	0.395688
	Foreground MAE ↓	0.083418	0.081792	0.084973

E.4 Uncertainty Estimates

Table 17 reports the episode-cluster bootstrap intervals for each model and scoring window; Table 18 gives the paired Khora-minus-ablation differences. The resampling protocol and scope of these intervals are described in Section E.1.

Table 17: **Continued.** Episode-cluster bootstrap intervals, frames 38–57.

Method	Metric	Estimate	Lower 95%	Upper 95%
Relative frames 38–57				
Gamma-World	Detection F1 ↑	0.019482	0.015037	0.024414
	Detection recall ↑	0.015064	0.011583	0.018869
	Direction accuracy (4 frames) ↑	0.092087	0.083380	0.101183
	Absolute pose PCK ($0.2h$) ↑	0.018057	0.013859	0.022148
	Foreground SSIM ↑	0.147564	0.142932	0.152811
	Foreground MAE ↓	0.226105	0.220287	0.231622
MIRA	Detection F1 ↑	0.452510	0.438689	0.467047
	Detection recall ↑	0.429884	0.416224	0.444601
	Direction accuracy (4 frames) ↑	0.450706	0.435808	0.466345
	Absolute pose PCK ($0.2h$) ↑	0.412928	0.395976	0.430564
	Foreground SSIM ↑	0.255510	0.250030	0.261062
	Foreground MAE ↓	0.126754	0.123871	0.129498
Khora w/o Cross-Agent State	Detection F1 ↑	0.008690	0.004995	0.012845
	Detection recall ↑	0.004656	0.002663	0.006895
	Direction accuracy (4 frames) ↑	0.004259	0.002759	0.005863
	Absolute pose PCK ($0.2h$) ↑	0.004485	0.002416	0.006921
	Foreground SSIM ↑	0.220065	0.215129	0.225481
	Foreground MAE ↓	0.162544	0.158687	0.166021
Khora	Detection F1 ↑	0.902400	0.895124	0.909580
	Detection recall ↑	0.864516	0.853689	0.875000
	Direction accuracy (4 frames) ↑	0.692939	0.677961	0.706945
	Absolute pose PCK ($0.2h$) ↑	0.738481	0.724728	0.752238
	Foreground SSIM ↑	0.360863	0.356268	0.365589
	Foreground MAE ↓	0.086796	0.085012	0.088439

Table 18: **Paired Khora minus ablation differences.** Signed differences in original metric units with episode-cluster bootstrap 95% intervals. Negative MAE differences favor Khora; positive differences favor Khora for the other metrics.

Frames	Metric	Difference	Lower 95%	Upper 95%
13–32	Detection F1 ↑	0.619872	0.607150	0.631774
	Detection recall ↑	0.687097	0.675364	0.698097
	Direction accuracy (4 frames) ↑	0.646717	0.631147	0.661243
	Absolute pose PCK ($0.2h$) ↑	0.571507	0.557994	0.585445
	Foreground SSIM ↑	0.142285	0.138408	0.146243
	Foreground MAE ↓	-0.058308	-0.060516	-0.056113
38–57	Detection F1 ↑	0.893710	0.885384	0.901474
	Detection recall ↑	0.859860	0.849057	0.870239
	Direction accuracy (4 frames) ↑	0.688680	0.673438	0.702869
	Absolute pose PCK ($0.2h$) ↑	0.733997	0.720478	0.747374
	Foreground SSIM ↑	0.140798	0.136622	0.145054
	Foreground MAE ↓	-0.075748	-0.078307	-0.073170

Table 19: **Detection-threshold sensitivity on all 1,643 cases.** Raw detections, GT boxes, and geometric filtering are fixed. Only the generated detection confidence threshold changes. P, R, and F1 pool the displayed counts; 0.25 reproduces the main evaluation. Ranks compare models at the same threshold and window. Baselines: Solaris (Savva et al., 2026), Gamma-World (Liu et al., 2026), and MIRA (Hu et al., 2026a).

Method	Conf.	TP	FP	FN	Precision ↑	Recall ↑	F1 ↑
Relative frames 13–32							
Solaris	0.15	5341	25917	27519	0.170868	0.162538	0.166599
	0.25	5295	24777	27565	0.176077	0.161138	0.168277
	0.35	5255	23966	27605	0.179836	0.159921	0.169295
	0.50	5135	22478	27725	0.185963	0.156269	0.169828
Gamma-World	0.15	1859	25219	31001	0.068654	0.056573	0.062031
	0.25	1724	21090	31136	0.075568	0.052465	0.061932
	0.35	1607	18486	31253	0.079978	0.048904	0.060695
	0.50	1439	15354	31421	0.085690	0.043792	0.057962
Khora w/o Cross-Agent State	0.15	6354	4918	26506	0.563698	0.193366	0.287954
	0.25	6079	3842	26781	0.612741	0.184997	0.284192
	0.35	5815	3145	27045	0.648996	0.176963	0.278097
	0.50	5346	2376	27514	0.692308	0.162690	0.263467
Khora	0.15	29261	2802	3599	**0.912610**	**0.890475**	**0.901406**
	0.25	28657	1879	4203	**0.938466**	**0.872094**	**0.904063**
	0.35	28102	1315	4758	**0.955298**	**0.855204**	**0.902484**
	0.50	26857	748	6003	**0.972903**	**0.817316**	**0.888349**

Table 19: **Continued.** Detection-threshold sensitivity, frames 38–57.

Method	Conf.	TP	FP	FN	Precision ↑	Recall ↑	F1 ↑
Relative frames 38–57							
Gamma-World	0.15	547	21626	32313	0.024670	0.016646	0.019879
	0.25	495	17461	32365	0.027567	0.015064	0.019482
	0.35	456	14978	32404	0.029545	0.013877	0.018884
	0.50	411	12230	32449	0.032513	0.012508	0.018066
MIRA	0.15	14459	16962	18401	0.460170	0.440018	0.449869
	0.25	14126	15448	18734	0.477649	0.429884	0.452510
	0.35	13806	14449	19054	0.488621	0.420146	0.451804
	0.50	13092	12948	19768	0.502765	0.398418	0.444550
Khora w/o Cross-Agent State	0.15	170	3083	32690	0.052259	0.005173	0.009415
	0.25	153	2200	32707	0.065023	0.004656	0.008690
	0.35	139	1677	32721	0.076542	0.004230	0.008017
	0.50	109	1157	32751	0.086098	0.003317	0.006388
Khora	0.15	29016	2654	3844	**0.916198**	**0.883019**	**0.899303**
	0.25	28408	1693	4452	**0.943756**	**0.864516**	**0.902400**
	0.35	27784	1199	5076	**0.958631**	**0.845526**	**0.898533**
	0.50	26419	728	6441	**0.973183**	**0.803987**	**0.880531**

E.5 Detection-Threshold Sensitivity

Table 19 gives the complete counts and detection scores at all four confidence thresholds, separately for the two scoring windows. Only the generated detection threshold varies; the GT references, cohort, and geometric filters remain fixed.

Table 20: **All measured compute-only scaling points.** Times are ms per four-frame step; memory is GiB per GPU. Throughput is view-frames/s, distributed over one GPU per agent.

Agents	STBoard ↓	Project ↓	Synthesis ↓	Total ↓	FPS ↑	GiB/GPU ↓	View-fps ↑
1	26.7359	4.8839	75.5396	107.1594	37.3276	8.1621	37.3276
2	26.7343	5.0808	75.6049	107.4200	37.2370	8.0117	74.4740
4	26.8281	4.9044	75.8272	107.5597	37.1886	7.9365	148.7545
8	26.9591	5.4382	76.5093	108.9066	36.7287	7.8989	293.8299
16	27.1625	6.0219	77.4040	110.5884	36.1702	7.8801	578.7225
32	27.6174	6.3001	76.9649	110.8824	36.0742	7.8707	1154.3757
48	28.0366	7.1105	78.2479	113.3950	35.2749	7.8676	1693.1964
64	28.9105	7.2967	78.8271	115.0343	34.7722	7.8660	2225.4233
80	29.3795	8.1325	79.2210	116.7330	34.2662	7.8651	2741.2984

Table 21: **Observed compute-only timing ranges.** Each entry is minimum–maximum in ms; these ranges are not confidence intervals and component extrema need not co-occur.

Agents	STBoard ↓	Projection ↓	Synthesis ↓	Total ↓
1	26.488–27.476	4.310– 8.740	75.040–77.740	105.838–113.956
2	26.506–27.172	4.310– 7.870	74.980–77.980	105.796–113.022
4	26.584–27.341	4.370– 6.610	75.150–79.600	106.104–113.551
8	26.692–27.439	4.600– 8.710	75.190–84.180	106.482–120.329
16	26.960–27.710	4.590–29.390	75.520–92.210	107.070–149.310
32	27.378–28.277	4.690–12.270	75.400–80.580	107.468–121.127
48	27.773–29.507	4.930–12.100	75.590–91.400	108.293–133.007
64	28.549–30.024	4.920–17.820	75.660–91.260	109.129–139.104
80	28.950–30.923	5.080–28.840	75.700–96.750	109.730–156.513

F Runtime and Scalability

F.1 Compute-Only Measurements

Figure 4 reports compute-only scaling measurements with populations 1, 2, 4, 8, 16, 32, 48, 64, and 80. One GPU is assigned per agent. Each rollout step emits four frames, so per-view FPS is $4000/L$ for step latency L in ms, and distributed throughput is $N \cdot 4000/L$. Compute-only latency sums STBoard update, projection, and view synthesis (model sampling plus VAE decoding). Image encoding, network transport, and client delivery are excluded.

F.2 Measurement Scope and Timing Ranges

We exclude deadline-fallback worker results from view-synthesis compute timing. Table 22 reports the mean and maximum fallback-worker counts per step. Per-GPU memory is the mean of measured GPU peaks. Logical total memory is calculated by multiplying this value by the active agent count.

Table 21 reports observed minimum and maximum timings. Projection accounts for 4.56–6.97% of mean compute-step latency across the measured populations.

F.3 End-to-End Delivery Performance

As the population increases from 1 to 80 agents, end-to-end latency grows from 113.5055 to 205.6232 ms; corresponding per-view throughput decreases from 35.2406 to 19.4531 FPS. In addition to model computation, these measurements include image encoding, network transport, and client delivery.

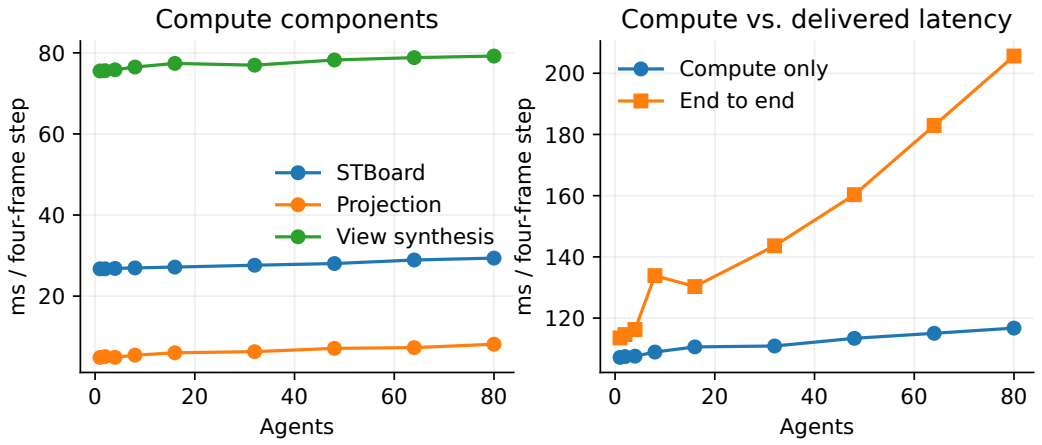

Figure 16: **Measured runtime components and delivery overhead.** Left: STBoard update, projection, and view-synthesis latency. Right: compute-only versus end-to-end latency, including delivery in the latter. Points show measured populations from the same experiment.

Table 22: **End-to-end latency and fallback counts.** Times are ms; FPS includes delivery. Fallback columns report the mean and maximum number of fallback workers per measured step.

Agents	Image transport ↓	Server ↓	End-to-end ↓	FPS ↑	Fallback mean ↓	Max ↓
1	2.4563	111.8901	113.5055	35.2406	0.0000	0
2	3.1265	112.5721	114.6192	34.8982	0.0000	0
4	4.0102	113.3196	116.2761	34.4009	0.0000	0
8	5.7167	129.0565	133.8820	29.8771	0.4100	1
16	10.3416	122.0269	130.3112	30.6957	0.0000	0
32	19.1814	128.4396	143.6130	27.8526	0.0000	0
48	29.2574	137.9087	160.3278	24.9489	0.0000	0
64	42.8465	150.2562	182.9297	21.8663	0.0000	0
80	55.4797	163.0364	205.6232	19.4531	1.2800	53

Table 23: **Compute-only projections, not measurements.** Values extend the measured linear latency fit. Logical memory and throughput assume one allocated GPU per agent.

Agents	Total ms ↓	FPS ↑	GiB/GPU ↓	Logical GiB ↓	View-fps ↑
128	122.5916	32.6287	7.8651	1006.7313	4176.4678
256	137.6671	29.0556	7.8651	2013.4625	7438.2348
512	167.8179	23.8354	7.8651	4026.9250	12203.7007
1024	228.1197	17.5347	7.8651	8053.8500	17955.4856

F.4 Extrapolation Beyond Measured Populations

We extrapolate compute-only latency to unmeasured populations of 128, 256, 512, and 1,024 agents using the fitted linear model $L(N) = 107.516194 + 0.117777N$ ms, with $R^2 = 0.975822$ on the measured populations. Table 23 lists every projected point used by the figure. Projections hold memory per GPU at the measured 80-agent value and assume one GPU per agent. Network costs are excluded.

G Additional Qualitative Results

G.1 Dynamic Population Examples

Figure 17 enlarges all four active views from the middle stage of Figure 5. Figure 18 enlarges the initial and final two-agent stages. Together, these panels show the $2 \rightarrow 4 \rightarrow 2$ population-change sequence at a larger scale, complementing the main figure rather than adding independent test cases.

G.2 Full-Frame Context for the Actor Comparison

Figures 19 and 20 show the complete scoring images for case 79 from the sampled inspection set. The early panel uses source frames 13, 22, and 32; the late panel uses 38, 47, and 57. The same ego view and physical frames are used within each panel, including the Solaris one-frame offset. These examples complement the enlarged actor crops rather than replacing full-cohort metrics. They are not selected to maximize any model's score.

These full-frame panels show the surrounding scene, first-person body parts, and any missing or misplaced players. Such context is useful when assessing whether a high scene-level score reflects the target character. Quantitative world and action scores remain distinct because their reference objects, regions, and temporal windows differ.



Agent 1, t_1



Agent 2, t_1



Agent 3, t_1



Agent 4, t_1

Figure 17: **Dynamic population: four-agent stage.** Enlarged views of all four active agents at t_1 in Figure 5. Agent numbering and stage labels are consistent with the main figure.



Agent 1, t_0



Agent 2, t_0



Agent 3, t_2



Agent 4, t_2

Figure 18: **Dynamic population: initial and final stages.** Top: the two initially active agents. Bottom: the two remaining agents after Agents 1 and 2 leave. Agent numbering follows Figure 5.



Figure 19: **Full-frame generated context, early window.** Case 79, source frames 13, 22, and 32 from top to bottom. Columns share the same GT camera view and physical time. Images use the scoring resize/padding protocol, with no learned enhancement, registration, or crop selection. The six columns keep the same model order in both windows; MIRA conditioning frames are explicitly marked rather than shown as generated results.

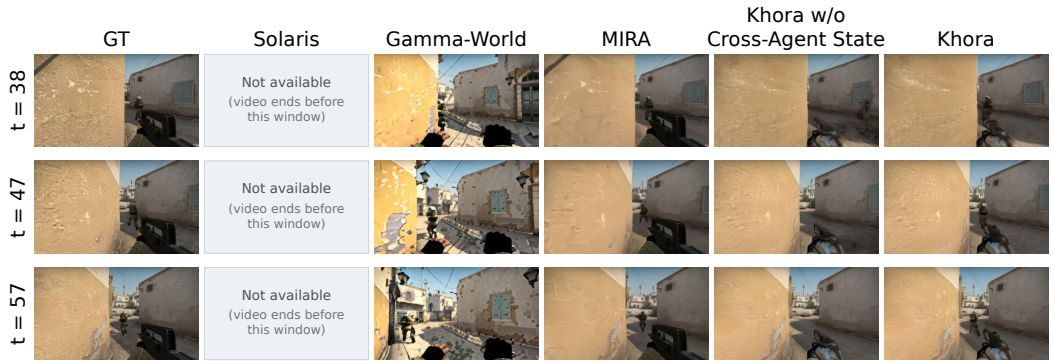

Figure 20: **Full-frame generated context, late window.** The same case and ego view at source frames 38, 47, and 57. MIRA is available in this window; Solaris is not. No single purely generated window exists across all five models, so early and late examples are shown separately. The Solaris column remains visible and is marked unavailable.