OASIS-WORLD: TOWARD OPEN-DOMAIN MULTI-AGENT INTERACTIVE WORLD MODEL

Anonymous authors
Paper under double-blind review



Figure 1: **Selected samples from OasisWorld.** The upper and lower panels show two distinct examples where the model takes initial frames from each agent as input to generate action-conditioned interactive video sequences.

ABSTRACT

World models capable of predicting environmental dynamics conditioned on agent actions form a cornerstone of embodied intelligence and interactive simulation. Although single-agent formulations have witnessed substantial progress, extending world modeling to multi-agent environments introduces fundamental challenges including identity persistence under severe mutual occlusion, dynamic spatial tracking across co-present entities, and the grounding of concurrent behaviors to distinct visual outcomes, all of which substantially exceed the capacity of existing datasets. To bridge this data bottleneck, we present **OasisData-Engine**, an automated generation pipeline that integrates scene preparation, coordinated trajectory planning, offline photorealistic rendering, and streaming annotation to facilitate scalable parallel capture of multi-agent interactions within high-fidelity Unreal Engine environments. Leveraging this infrastructure, we establish **OasisData**, a comprehensive benchmark encompassing over 130 environments and 80 distinct character identities with configurations supporting from one to ten concurrent participants, accumulating more than 4,900 hours of video strictly paired with fine-grained frame-level action commands and structured descriptions. Furthermore, we develop **Oasis-World** as a baseline for open-domain multi-agent world modeling, conducting systematic empirical evaluations across spatial reasoning, action responsiveness, and occluded interactions to delineate critical challenges and future trajectories for interactive world modeling. **We will open source the model weights, code and data**.

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

1 INTRODUCTION

Recent advances in the visual fidelity and temporal coherence of video generation have spurred the development of interactive video world models (Yu et al., 2025b; Wang et al., 2026c; Huang et al., 2026a; Wu et al., 2026b; Team et al., 2026a). Using action commands as control inputs, these models predict future visual observations from an agent's observation and action history (Yu et al., 2025a; 2026; Huang et al., 2025; 2026b; Wei et al., 2026; Xiao et al., 2026), with promising applications in synthetic data generation, online planning, and policy evaluation for embodied intelligence (Qin et al., 2024; Jeon et al., 2026; Zhao et al., 2026; Wang et al., 2026b; Yao et al., 2026). However, most existing world models simulate only a single agent's viewpoint and responses to its actions. In real-world environments, multiple agents act concurrently and jointly shape how a scene evolves, motivating multi-agent video world models that capture their behaviors and mutual influences.

Compared with single-agent settings, multiple agents sharing the same space introduce substantially more complex relative motion and mutual occlusion. Models must preserve each character's identity as characters move past one another, while associating each agent's control inputs with the correct character motions and local visual changes. Existing work on multi-agent world models has largely focused on specific game environments, such as Minecraft and It Takes Two (Savva et al., 2026; Wu et al., 2026a; Liu et al., 2026; Cai et al., 2026; Hu et al., 2026). Reliance on these closed environments poses challenges for transfer to complex real-world settings. Developing open-domain multi-agent world models therefore requires training data that combine precise interaction supervision with broad content diversity. Such data should comprehensively capture concurrent action trajectories and dynamic occlusions across agents to supervise the correspondence between individual action commands and local visual changes. They should also span diverse natural environments and character appearances to support generalization to open environments.

To this end, we develop **OasisData-Engine**. The system integrates scene preparation, planning side, and rendering side to enable automated, parallel data collection in high-fidelity 3D environments built with Unreal Engine, along with synchronized records of actions and spatial positions.

Using this engine, we construct **OasisData**, a large-scale action-conditioned video dataset for open-domain multi-agent world models. It covers more than 130 scenes and over 80 character types, with interaction settings ranging from a single agent to 10 simultaneously active agents and over 4,900 hours of video. To support fine-grained world model training, the dataset provides dense frame-level action signals and multidimensional structured annotations. By aligning high-resolution multi-agent interaction videos with environmental semantics and precise control actions, **OasisData** supports research on visual scene evolution, spatial interactions, and responses to concurrent actions in multi-agent settings.

We further introduce **Oasis-World** as a baseline for studying multi-agent world models using **OasisData**. Our analysis examines character positions, action responses, and interactions under mutual occlusion. Our contributions can be summarized as follows:

- We develop **OasisData-Engine**, which integrates scene preparation, trajectory planning, and offline rendering to enable automated, parallel data collection in high-fidelity 3D environments built with Unreal Engine, with synchronized structured annotations.
- We introduce **OasisData**, a dataset with 130+ scenes, 80+ characters, and settings ranging from a single agent to 10 simultaneously active agents. It contains over 4,900 hours of video with structured annotations and frame-level action labels.
- We propose **Oasis-World** as a baseline for open-domain multi-agent world modeling and analyze character positions, action responses, and interactions under mutual occlusion.

2 RELATED WORK

2.1 VIDEO DATASETS FOR WORLD MODELING

Large-scale video-text corpora provide foundational pre-training data for generative backbones. Milestones such as WebVid-10M (Bain et al., 2021), Panda-70M (Chen et al., 2024), and InternVid (Wang et al., 2022) offer rich visual diversity and semantic associations, advancing open-domain video

Table 1: **Comprehensive comparison of recent multi-modal and embodied video datasets.** We compare domain coverage, perspective, multi-agent support, action granularity, captioning, spatial grounding, and total scale. "–" denotes undisclosed or publicly unavailable information.

Dataset	Domain	1st-P (Ego)	3rd-P (Exo)	Multi-Agent	Action Granularity	Caption Type	Spatial Grounding	Scale / Video Clips
ABot-World (Jiang et al., 2026)	Open domain	✓	✓	✗	Frame-level	Dense caption	Estimated	30,969 clips
MIND (Ye et al., 2026)	Open domain	✓	✓	✗	Frame-level	Dense caption	Ground-truth	250 clips
MiraData (Ju et al., 2024)	Open domain	✓	✓	✗	Textual	Structured caption	Textual relation	788,000 clips
OmniWorld (Zhou et al., 2025)	Open domain	✓	✓	✗	Trajectory	Dense caption	Estimated	600,000+ clips
WildWorld (Li et al., 2026b)	Monster Hunter Wilds	✗	✓	✗	Frame-level	Structured caption	Ground-truth	3,434 clips
Sekai (Li et al., 2026a)	Open domain	✓	✗	✗	Trajectory	–	Estimated	325,635 clips
SpatialVID (Wang et al., 2026a)	Open scene	✓	✗	✗	Trajectory	Structured caption	Estimated	2,710,000 clips
Solaris (Savva et al., 2026)	Minecraft	✓	✗	2 agents	Frame-level	–	Ground-truth	9,240 clips
MultiWorld (Wu et al., 2026a)	It Takes Two + Robot Sim	✗	✓	2 agents	Frame-level	–	Ground-truth	Game: ∼100 h Sim: 12,000 eps
Gamma-World (Liu et al., 2026)	Minecraft + Bimanual Robot	✓	✗	2 / 4 agents	Frame-level	–	–	–
OasisData	**Open domain**	✓	✓	**1–10 agents**	**Frame-level**	**Structured caption**	Ground-truth	**297,582 clips (4,959h)**

generation. However, supervision in these datasets relies solely on superficial captions without explicit action signals or structured metadata, failing to support action-conditioned world modeling.

To introduce spatial awareness, subsequent works curate geometric video datasets. Benchmarks like RealEstate10K (Zhou et al., 2018) and ScanNet (Dai et al., 2017) provide camera trajectories with dense depth, while OmniWorld (Zhou et al., 2025) aggregates cross-domain 4D assets for novel-view synthesis. Although these benchmarks enhance spatial consistency, they largely focus on static or low-dynamic scenes without control signals synchronized with character behaviors, limiting their utility for dynamic world simulation.

Recent interactive world models incorporate explicit action signals. Specifically, ABot-World (Jiang et al., 2026) models gameplay sequences, SpatialVID (Wang et al., 2026a) curates interactive exploration trajectories with structured captions, and MIND (Ye et al., 2026) leverages Unreal Engine to couple ground-truth camera poses with temporal contexts. While facilitating action-conditioned generation and embodied policy learning, these datasets remain confined to single-agent dynamics in passive surroundings. Non-agent entities are treated as static clutter, failing to capture inter-agent spatial relations, reciprocal behaviors, and identity preservation essential for collective dynamics.

To transcend these domain constraints, we introduce **OasisData-Engine**, an automated pipeline for parallel data collection in photorealistic Unreal Engine environments with real-time streaming of synchronized action labels and spatial coordinates. Powered by this engine, we construct **OasisData**, a large-scale benchmark for multi-agent world modeling. Featuring over 130 environments and 80 character assets, OasisData spans dynamic configurations from 1 to 10 interacting agents across more than 4,900 hours, establishing a foundational data infrastructure for scalable multi-agent research.

2.2 Interactive World Models

Interactive world models aim to forecast the future dynamic evolution of an environment conditioned on historical visual observations and control signals. Early explorations concentrated predominantly on specific domains, exemplified by UniSim (Yang et al., 2023) and GAIA-1 (Hu et al., 2023) for robotic control and autonomous driving, alongside GameGen-X (Che et al., 2025) and MineWorld (Guo et al., 2025) tailored for gaming scenarios. Building upon these efforts, subsequent frameworks including Matrix-game 2.0 (He et al., 2025), Infinite-World (Wu et al., 2026b), and the Lingbot-World series (Team et al., 2026b) have further advanced the interactive generation paradigm pioneered by Genie (Bruce et al., 2024), substantially improving rollout horizons, action responsiveness, and real-time inference efficiency in single-agent environments.

Nevertheless, real-world physical environments and complex virtual worlds are fundamentally driven by the co-presence and mutual coordination of multiple entities, pushing world modeling toward multi-agent scenarios. Relative to single-agent settings, multi-agent formulations introduce inherent challenges, including action-and-reaction interactions among participants, cross-view spatiotemporal consistency, and long-horizon identity preservation.

To address these challenges, Solaris (Savva et al., 2026) proposed a dedicated multi-view Diffusion Transformer architecture, though it was only validated on Minecraft. Meanwhile, Gamma-World (Liu

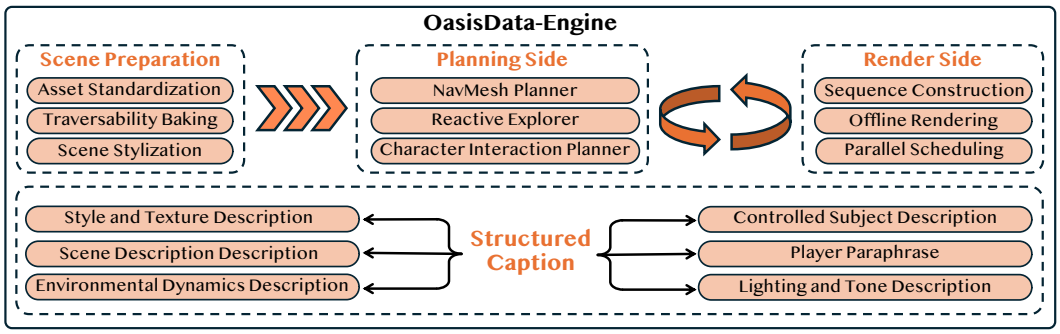

Figure 2: **OasisData-Engine overview.** Scene preparation runs once per environment. The planning and rendering sides then operate in close coordination to generate multi-agent trajectories and corresponding videos, finally leveraging VLMs to automatically produce structured descriptions.

et al., 2026) devised Simplex Rotary Agent Encoding combined with sparse Hub attention, achieving linear computational complexity across agents while preserving permutation equivariance, yet its evaluation was likewise confined to Minecraft and robotic manipulation environments. Because their representations remain confined to these specific domains, existing frameworks struggle to generalize to photorealistic open-domain multi-agent interactions. To bridge this gap, we propose **Oasis-World**, an open-domain multi-agent world model baseline built upon **OasisData** to investigate collective world modeling across diverse scenes and heterogeneous interaction patterns.

3 OasisData-Engine

As illustrated in Figure 2, OasisData-Engine establishes an automated pipeline tailored for scalable multi-agent interactive video generation, decoupling the entire synthesis procedure into four stages. First, scene preparation, detailed in Section 3.1, standardizes raw heterogeneous assets and bakes navigation topologies to provide structured geometric substrates with controllable stylistic appearances. Building upon these spatial bounds, the planning pipeline elaborated in Section 3.2 synthesizes collision-free multi-agent trajectories that reconcile broad spatial exploration with intentional spatiotemporal encounters. The rendering module introduced in Section 3.3 subsequently translates the planned kinematics into sequence tracks to execute high-fidelity multi-view offline rendering under concurrent GPU scheduling. Finally, the structured annotation stage presented in Section 3.4 enriches the synthesized video frames with disentangled environmental dynamics and multi-agent visual attributes to deliver grounded semantic metadata.

3.1 Scene Preparation

Scene preparation converts raw 3D assets into structured, simulation-ready environments for subsequent planning and rendering. It sequentially standardizes heterogeneous asset formats, bakes collision-aware navigation meshes for topological constraints, and configures material and lighting presets to ensure controllable visual diversity.

Asset Standardization. Asset packages from disparate Unreal Engine releases are migrated uniformly to Unreal Engine 5.5, with streaming sub-levels and World Partition cells merged to yield standardized scenes that support headless loading. The rendering module can directly mount these consolidated environments for offline capture without initiating interactive editor sessions.

Traversability Baking. Navigation meshes are baked over walkable surfaces across the standardized environments to encode traversable geometric structures. The resulting navigation topologies are persisted alongside the level assets, providing explicit navigable bounds and physical collision geometry to support subsequent path planning algorithms.

Scene Stylization. We encapsulate each aesthetic configuration into an independent map variant integrated with calibrated lighting and Niagara weather systems, loading visual presets directly without runtime blueprint dependencies. In addition, we lock offline-calibrated exposure values and disable dynamic auto-exposure, preventing unintended photometric amplification in shadow regions to preserve accurate environmental contrast.

3.2 PLANNING SIDE

Trajectory planning synthesizes physically plausible and interactive multi-agent paths. To balance environmental navigation, collision avoidance, and inter-agent encounters, the framework unifies NavMesh pathfinding, reactive exploration, and character interaction. Furthermore, stochastic waypoint sampling diversifies these motion sequences to mitigate behavioral homogenization.

NavMesh Planner. The NavMesh planner optimizes for spatial exploration and collision avoidance. Agents initialize from randomly chosen reachable locations and traverse along pre-baked navigation meshes toward sampled target destinations while incorporating occasional heading perturbations. The resulting paths conform strictly to traversable surfaces, providing kinematically stable baseline trajectories for subsequent encounter scheduling.

Reactive Explorer. The reactive explorer detects local obstacles through a fan-shaped array of short-range ray casts, prioritizing unvisited directions to reconcile obstacle avoidance with forward exploration. Because this strategy operates independently of pre-baked navigation topologies, its local decision-making may occasionally induce wandering or backtracking patterns, thereby serving as an auxiliary mechanism alongside the NavMesh planner to expand overall spatial coverage.

Character Interaction Planner. The character interaction planner introduces spatiotemporal rendezvous constraints onto the baseline paths synthesized by the preceding strategies. By modulating waypoint timing profiles and orientation angles, it orchestrates multiple agents to converge within designated spatial areas. This explicit scheduling organizes close-range collective behaviors, supporting the acquisition of rich multi-view dynamic interaction observations.

3.3 RENDER SIDE

The rendering side converts planned multi-agent trajectories into multi-view video corpora. It binds camera paths and character kinematics into Level Sequences, renders photorealistic frames offline, and post-processes them into standardized video streams. Meanwhile, a parallel scheduler coordinates concurrent rendering workers with instance reuse to sustain high data throughput.

Offline Rendering. We batch-render video sequences using Unreal Engine's Movie Render Queue. To eliminate aliasing artifacts, the pipeline accumulates spatiotemporal subsamples for each individual frame. Rather than using static lightmaps, we employ Lumen dynamic global illumination to compute multi-bounce indirect lighting, which instantly updates radiometric calculations to moving viewpoints and dynamic interactions while preserving photorealistic visual consistency.

Sequence Construction. We import planned camera and character trajectories into Unreal Engine 5, serializing spatial transforms into Level Sequence tracks along a unified timeline. The Movie Render Queue renders viewpoint frames governed by the Camera Cut track into HDR EXR sequences. Concurrently, worker threads perform color-space transformation and tone mapping, immediately purging raw intermediate buffers to restrict disk storage footprints.

Parallel Scheduling. To maximize throughput under VRAM limits, multiple headless engine instances run concurrently per GPU, offloading video encoding to hardware encoders while dynamically scaling CPU post-processing threads. We persistently reuse active instances across sequences to amortize the overhead of engine initialization, shader compilation, and illumination warm-up, supported by an external watchdog that automatically reclaims stalled processes.

3.4 STRUCTURED CAPTION

We employ perspective-tailored prompts with Gemini-3.5-Flash to generate six-dimensional structured captions, disentangling global environmental attributes from fine-grained agent interactions across first-person and third-person views. At the environment level, four dedicated fields systematically record visual style and surface textures, spatial asset layout relative to the camera, ambient physical dynamics, and illumination atmosphere. At the agent level, the description details the primary actor's kinetic postures, holistic appearances, and facial features. Furthermore, a peer paraphrase field characterizes this actor from the external vantage points of co-present agents, establishing explicit semantic priors to enforce cross-view consistency and identity alignment across parallel observation streams. The complete prompt designs and concrete templates for both viewpoints are documented in Appendix D.

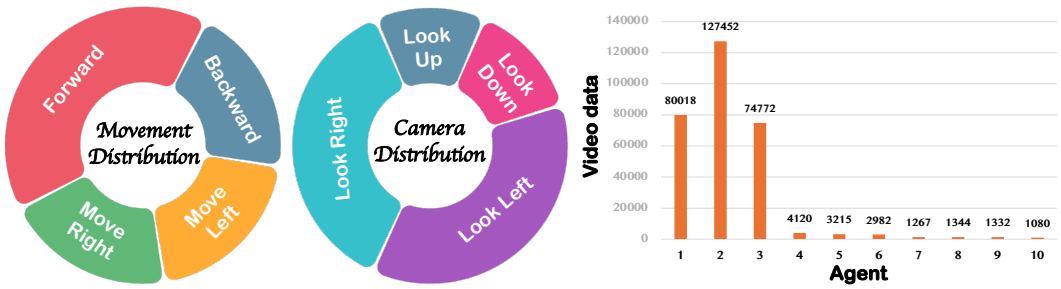

Figure 3: **Statistical information of OasisData.** The left and middle charts illustrate the balanced distribution of agent movement directions and camera rotation actions respectively. The right histogram shows the breakdown of video data volume across scenes with varying numbers of agents.

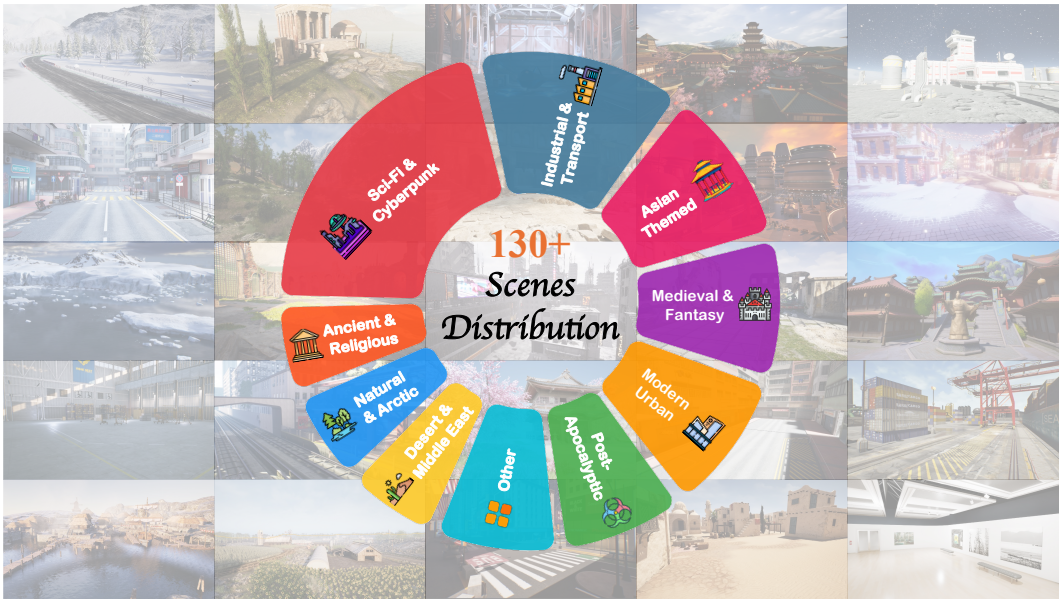

Figure 4: **Scene distribution in OasisData.** The dataset covers over 130 environments across diverse themes such as urban, sci-fi, and natural landscapes. The central chart depicts the thematic breakdown, accompanied by representative visual examples in the background.

4 OASISDATA

As illustrated in Figure 3, **OasisData** comprises 297,582 video sequences curated across 132 diverse environments, accumulating over 428.5 million frames and 4,959 hours of total footage. The benchmark systematically scales from single-agent dynamics to multi-agent settings with up to 10 concurrent participants. To capture realistic interaction dynamics, the distribution places primary emphasis on dense local interactions among 1 to 3 agents, while preserving a steady tail that spans up to 10 entities to study broader group choreography. Furthermore, the underlying action space exhibits balanced coverage across omnidirectional planar locomotion and decoupled camera rotations in both pitch and yaw. This orthogonal coverage prevents trivial straight-line path bias, driving thorough three-dimensional spatial exploration and providing dense behavioral supervision across diverse collaborative scenarios.

4.1 SCENE

The **OasisData** scene library encompasses 132 photorealistic Unreal Engine environments, spanning spatial scales from compact single rooms to sprawling city districts. Built with production-grade materials and rich geometric details, these environments fall into four balanced semantic domains illustrated in Figure 4, comprising Historical & Fantasy, Modern & Industrial, Sci-Fi & Futuristic, and Natural & Wilderness. Each domain accounts for roughly a quarter of the corpus, ensuring broad visual and topological diversity without category bias.



Figure 5: **Character distribution in OasisData.** The dataset features over 80 distinct characters across diverse archetypes. The central chart shows the category breakdown, accompanied by representative character renders in the background.

4.2 Character

Character diversity is essential for robust world modeling, governing how reliably generative models generalize across varied morphologies and visual appearances. **OasisData** incorporates over 80 distinct characters, each calibrated to realistic physical scales and enforced with foot-ground contact constraints to ensure kinematic plausibility. As illustrated in Figure 5, these assets span diverse functional and cultural archetypes ranging from tactical combatants to everyday civilians, providing comprehensive motion and appearance priors for multi-agent interaction learning.

5 Oasis-World

Oasis-World is an open-domain multi-agent world model baseline trained on **OasisData** to explore action-conditioned video prediction across single-agent and multi-agent settings. This section formalizes the problem and then presents the key architectural designs.

5.1 Preliminaries

Oasis-World generates synchronized visual observation sequences across multiple interacting agents conditioned directly on their action inputs. At each time step t, the visual observations and control signals across all P agents are grouped into joint states $x_t = \{x_t^1, \ldots, x_t^P\}$ and joint actions $a_t = \{a_t^1, \ldots, a_t^P\}$. Stacking these representations over a temporal horizon of length T yields a joint latent observation tensor $x \in \mathbb{R}^{B\times P\times T\times H\times W\times C}$ and an action tensor $a \in \mathbb{R}^{B\times P\times T\times D}$, where B is the batch size, (H, W, C) represent the latent spatial dimensions and channel depth, and D denotes the action feature dimension. We optimize Oasis-World using Conditional Flow Matching (Liu et al., 2022; Lipman et al., 2022). For clean latent video x and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we define the probability path $x_\sigma = (1-\sigma)x + \sigma\epsilon$ with velocity target $\epsilon - x$. The transformer network v_θ learns to predict the corresponding vector field via:

$$\mathcal{L}_\theta = \mathbb{E}_{x,\, a,\, \sigma,\, \epsilon}\left[\left\|v_\theta(x_\sigma, \boldsymbol{\sigma}, a) - (\epsilon - x)\right\|_2^2\right]. \tag{1}$$

To ensure cross-agent spatial-temporal synchronization, we enforce a unified noise level across all entities by sampling a global scale $\sigma \sim \mathcal{U}(0, 1)$ and broadcasting it along both agent and temporal dimensions via $\boldsymbol{\sigma} = \sigma \mathbf{1}_{P\times T}$, where $\mathbf{1}_{P\times T}$ is an all-ones matrix.

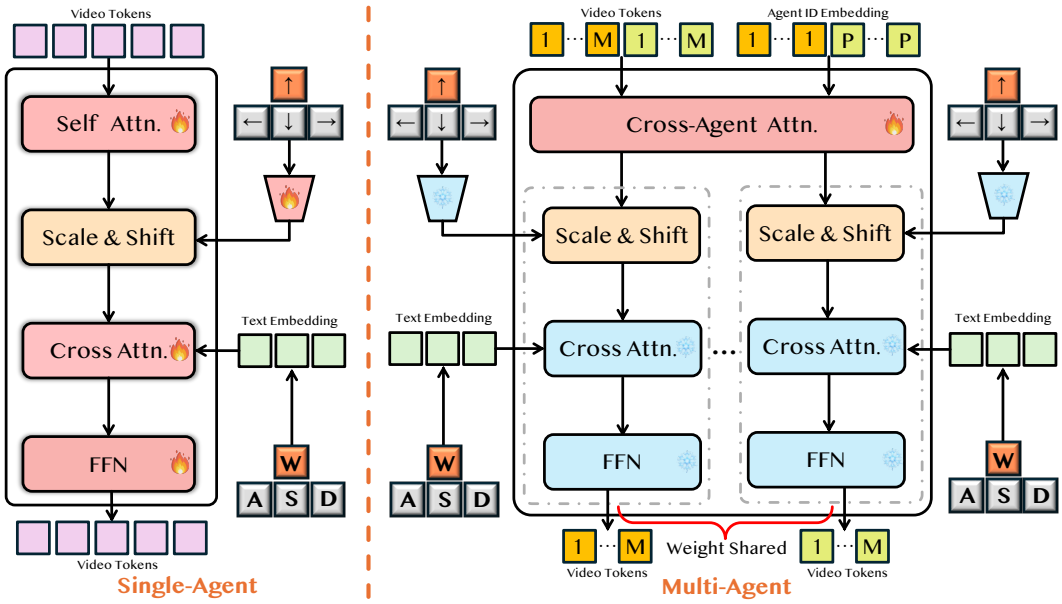

Figure 6: **Overview of Oasis-World. Left (Single-agent):** Decoupled action conditioning injects camera rotations via Scale & Shift and character movements via text-based Cross-Attention. **Right (Multi-agent):** Agent video tokens with Agent ID embeddings are jointly contextualized through Cross-Agent Attention, then processed by frozen, parameter-shared downstream layers. Flame and snowflake denote trainable and frozen parameters, respectively.

5.2 Network Architecture

Oasis-World builds upon Wan2.2-5B (Wan et al., 2025). We first continue pre-training the backbone on large-scale game data for single-agent action controllability, and subsequently extend it to multi-agent interactive settings. While OasisData supports up to ten agents, we focus our baseline on two-agent interactions to capture core reciprocal dynamics and mutual occlusions within a tractable context length. Figure 6 illustrates the DiT block designs and structural transition across both stages, with full training details deferred to Appendix B.

Single-Agent Backbone. Video frames are encoded by a frozen 3D VAE into latents for conditional denoising via a DiT, where each block sequentially applies Self-Attention, Scale & Shift modulation, Cross-Attention, and an FFN.

Decoupled Action Conditioning. We decouple control signals into dual pathways: low-level camera rotations are encoded via a 1D CNN and injected through scale-and-shift modulation, while high-level character locomotion is mapped to text embeddings for cross-attention injection, enabling orthogonal perspective and agent control.

Cross-Agent Attention. To support P interacting agents with M tokens each, we concatenate all video tokens into a unified joint context of size $P \times M$. Each agent token is augmented with an agent-specific identity embedding alongside 3D RoPE for spatiotemporal positional encoding. A Cross-Agent Attention layer then computes full attention over the concatenated sequence to exchange cross-view context, and splits the updated tokens back to their respective agent branches.

Parameter Sharing and Training Strategy. During multi-agent training, all downstream modules, including Scale & Shift, Cross-Attention, and the FFN, process individual agent streams independently while sharing identical weights. To preserve single-agent generative priors while minimizing computational overhead, we freeze all pre-trained single-agent parameters and exclusively optimize the newly introduced Cross-Agent Attention layers.

5.3 Discussion and Insights

To evaluate the zero-shot OOD robustness of **Oasis-World**, we construct an evaluation suite of 100 two-agent scenarios synthesized via GPT-Image-2. Human evaluators score the rollouts using a binary

Table 2: **Zero-shot OOD human evaluation on 100 synthesized two-agent scenarios.** We report the failure occurrence rate (%, lower is better) across four core multi-agent interaction challenges. Scored via binary assessment (0/1).

Method	Motion Divergence ↓	Position Drift ↓	Spatial Conflict ↓	Identity Loss ↓
Single-Agent Cascade	92% ± 2%	96% ± 3%	88% ± 8%	100% ± 1%
Oasis-World (Ours)	**67% ± 3%** (-25%)	**72% ± 2%** (-24%)	**76% ± 7%** (-12%)	**92% ± 1%** (-8%)



Figure 7: **Qualitative insights in Oasis-World.** We categorize the fundamental difficulties of multi-agent world modeling into four core dimensions, comprising cross-agent motion modeling, agent position tracking, cross-agent spatial consistency, and character preservation.

failure metric (0/1). As summarized in Table 2, Oasis-World with Cross-Agent Attention significantly outperforms the single-agent cascaded baseline across all metrics, validating the necessity of joint modeling. Furthermore, the results reveal four primary challenges:

Cross-Agent Motion Modeling. Current world models struggle to faithfully reflect the relative movement direction of other participants from an observer's perspective. When an agent traverses the environment, its heading direction and dynamic orientation perceived by other agents frequently diverge from the actual action commands, impairing coherent collective interactions.

Agent Position Drift. Models fail to maintain cumulative spatial coordinates under continuous locomotion. Rendered positions often lag or drift from planned trajectories, accumulating spatial errors that trigger abrupt coordinate snapping or visual teleportation.

Cross-Agent Spatial Consistency. Synthesizing mutually consistent environmental geometry remains a fundamental hurdle across distinct viewpoints. When multiple agents simultaneously inspect the same physical scene, the background layouts, static landmarks, and lighting conditions often exhibit severe cross-view contradictions.

Identity Preservation. Models frequently suffer from catastrophic character disappearance during close-range interactions and mutual occlusions. When one character walks in front of another and visual overlap occurs, the partially occluded subject often permanently vanishes in subsequent frames rather than properly recovering after disocclusion.

6 Limitations and Conclusion

We introduce OasisData-Engine to automate Unreal Engine data collection and construct OasisData. Spanning over 130 environments and 80 character types, the resulting 4,900 hours open-domain dataset provides frame-level actions and structured caption for up to ten agents. We also propose Oasis-World as an open-domain multi-agent world model baseline, highlighting core challenges in motion coordination, spatial consistency, and dynamic occlusion. While Oasis-World currently lacks explicit 3D geometry and persistent spatial memory for long-horizon consistency, it clearly demonstrates the feasibility of interactive multi-agent video generation and provides a solid foundation for interactive world model.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code development and data processing. Specifically, GPT-Image-2 generated the initial observation frames for the zero-shot OOD evaluation suite in Section 5.3, and Gemini-3.5-Flash was employed to annotate the video dataset with perspective-tailored structured captions in Section 3.4 and Appendix D. Additionally, we used generative AI tools to polish the language of the manuscript. We have thoroughly reviewed all AI-assisted work and take full responsibility for the final content of this paper, including code, data, text, results, and claims.

REPRODUCIBILITY STATEMENT

Section 5 describes the model architecture and training objectives; Appendix A specifies the detailed dataset curation and statistical characteristics; Appendix B reports the training procedures and comprehensive hyperparameter settings; Appendix C outlines the evaluation protocol and baseline model selection; and Appendix D provides the perspective-tailored prompt templates. As stated in the abstract, we will release the model weights, source code, and dataset to support replication and further study.

ETHICS STATEMENT

The synthetic dataset **OasisData** introduced in this work was entirely generated using Unreal Engine 5 (UE5). The authors retain full intellectual property rights to the assets and generated data, with no copyright infringements or conflicts of interest. Furthermore, the dataset contains no personally identifiable information or private data. Additionally, we conducted a human study to evaluate the interactive video rollouts in Appendix C. All participants provided formal informed consent prior to the study, their evaluation ratings were collected strictly anonymously and are reported exclusively in aggregate. This work complies with standard institutional research ethics and integrity guidelines.

REFERENCES

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

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

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

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

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

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

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

Anthony Hu, Václav Volhejn, Adrien Ramanana Rahary, Chris Mulder, Aditya Makkar, Alyx Liao, Amélie Royer, Manu Orsini, Adam Jelley, Eloi Alonso, et al. Multiplayer interactive world models with representation autoencoders. *arXiv preprint arXiv:2607.05352*, 2026.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, et al. Solarwm: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026a.

Junchao Huang, Ziyang Ye, Xinting Hu, Tianyu He, Guiyu Zhang, Shaoshuai Shi, Jiang Bian, and Li Jiang. Live: Long-horizon interactive video world modeling. *arXiv preprint arXiv:2602.03747*, 2026b.

Byeongguk Jeon, Seonghyeon Ye, JaeHyeok Doo, Sungdong Kim, Minjoon Seo, Hyungmok Son, and Kimin Lee. Roboworld: Fast and reliable neural simulators for generalist robot policy evaluation. *arXiv preprint arXiv:2607.01060*, 2026.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026a.

Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Wildworld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative arpg. *arXiv preprint arXiv:2603.23497*, 2026b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ringattention with blockwise transformers for near-infinite context. In *International Conference on Learning Representations*, volume 2024, pp. 3992–4008, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

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.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, et al. Alayaworld: Long-horizon and playable video world generation. *arXiv preprint arXiv:2607.06291*, 2026a.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. *arXiv preprint arXiv:2603.08546*, 2026b.

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

Zhengxuan Wei, Xu Guo, Xinghui Li, Xunzhi Xiang, Min Wei, Yiran Zhu, Qiulin Wang, Xintao Wang, Pengfei Wan, Xiangwang Hou, et al. Geometry-aware implicit memory for video world models. *arXiv preprint arXiv:2606.02436*, 2026.

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

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026b.

Junyuan Xiao, Dingkang Liang, Xin Zhou, Yixuan Ye, Tongtong Su, Guangmo Yi, Bin Xia, Qiang Lyu, Shurui Shi, Jun Huang, et al. Divide and conquer: Decoupled representation alignment for multimodal world models. *arXiv preprint arXiv:2605.01896*, 2026.

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

Nanjie Yao, Hao Wang, Chong Cheng, Zhikang Chen, Wenzhe Li, Jiafei Lyu, Li Shen, Peilin Zhao, Zongqing Lu, Gao Huang, et al. World models for embodied intelligence: From plausible to controllable to actionable. *arXiv preprint arXiv:2609.16697*, 2026.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. Mind: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

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

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

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models. *arXiv preprint arXiv:2606.31734*, 2026.

Rui Zhao, Kaiming Yang, Jifeng Zhu, Siyang Chen, Ziqi Wang, Weijia Wu, Kevin Qinghong Lin, Heng Wang, and Mike Zheng Shou. Dream. exe: Can video generation models dream executable robot manipulation? *arXiv preprint arXiv:2606.04811*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

APPENDIX OVERVIEW

This appendix is organized into four sections to complement the main manuscript. Section A provides comprehensive details on the **OasisData**. Section B presents the implementation specifics required to reproduce **Oasis-World**, detailing the model architecture, full training hyperparameters, and architectural scalability alongside inference optimization when scaling beyond two participants. Section C provides a comprehensive overview of the evaluation protocol, encompassing benchmark metrics, baseline model selection with exclusion rationales, and the standardized human assessment procedure. Finally, Section D documents the perspective-tailored prompt designs employed for six-dimensional structured caption extraction.

A OASISDATA DETAILS

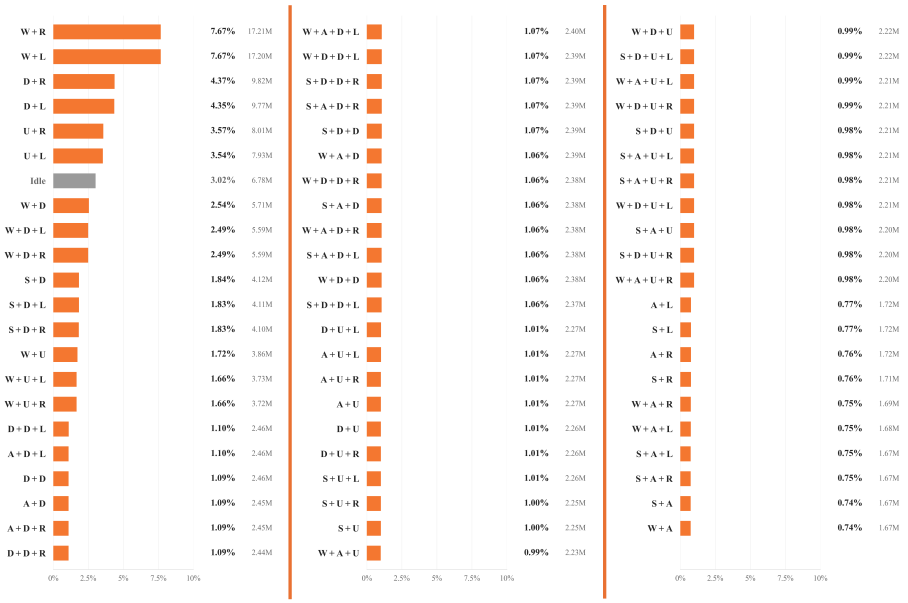

Figure 8: **Action combination distribution in OasisData.** We report the percentage (%) and total frame count (millions, M) across diverse concurrent control inputs, spanning agent locomotion (e.g., `W`, `A`, `S`, `D`), camera orientation adjustments (`U`, `D`, `L`, `R` denoting Up, Down, Left, and Right), and the `Idle` state. The smooth, well-balanced distribution across multi-key combinations ensures rich spatio-temporal exploration while preventing policy degeneration into trivial straight-line trajectories.

Data Scale and Perspective Allocation. OASISDATA covers ten distinct participant configurations ranging from single-agent setups to ten concurrent agents. Each configuration provides synchronized observations from both first-person egocentric viewpoints and third-person exocentric viewpoints. The dataset comprises 297,582 video sequences in total, divided into 203,149 first-person clips, representing 68.27%, and 94,433 third-person clips, accounting for the remaining 31.73%.

This deliberate skew toward first-person recordings addresses an inherent physical constraint in visual observability. Under an egocentric perspective, the camera view frustum is strictly bound to the agent's forward heading, making the probability of observing a peer entity's motion substantially lower than under an expansive third-person camera. Consequently, we deliberately prioritize first-person sequences to compensate for this perspective-induced observational sparsity and guarantee sufficient visual supervision for cross-agent interactive dynamics.

Action Space Parameterization. The control signals are parameterized as a four-dimensional discrete action vector governing longitudinal locomotion via keys W and S, lateral strafing via keys A and D, camera pitch via keys U and D, and camera yaw via keys L and R. Each control axis admits three orthogonal states, namely negative, neutral, and positive, yielding a complete combinatorial action space of 81 discrete states including the idle state where all controls remain

inactive. To eliminate syntactic ambiguity between identical letter notations, such as differentiating lateral right movement from pitch down rotation, all statistics are parsed directly on the four-dimensional numerical coordinates.

Across the entire benchmark, we record 509,771,264 valid frame-level action instances. As shown in Figure 8, the idle state accounts for merely 1.33%, single-axis activations represent 55.98%, and compound activations with two or more axes engaged simultaneously constitute 42.69%. This substantial proportion of concurrent control inputs provides dense supervision for learning complex trajectories and dynamic multi-agent coordination.

Table 3: **Training and model hyperparameters of single-agent stage.**

Setting	Value
Model Architecture and Video Setup	
Pretrained backbone	Wan2.2-5B
Video resolution ($H \times W$)	480×832
Video length / frame rate	121 frames / 24 FPS
Trainable parameters	DiT backbone
Optimization and Learning Schedule	
Optimizer	AdamW (`fused=True`)
Base learning rate	5×10^{-5}
Weight decay	0.01
Warm-up and LR schedule	ConstantLR (factor $1/3$, 5 warm-up steps), then constant
Gradient clipping	None
Numerical precision	BF16
Parallelism and Batch Configuration	
Global batch size	32 samples
Micro-batch size per DP replica	1
Gradient accumulation steps	1
Gradient checkpointing	Enabled
Data and Training Budget	
Training steps	30,000

B Implementation Details

B.1 Training Details and Hyperparameters

Oasis-World proceeds through two consecutive stages, namely single-agent training followed by multi-agent adaptation. The model first learns video generation guided by a single control stream, and is subsequently adapted to handle multiple concurrent control inputs and multiple interacting entities. The hyperparameters for both stages are listed in Table 3 and Table 4.

Single-Agent Stage. The model is initialized from the pretrained Wan2.2-5B backbone. In this stage, all parameters in the DiT backbone are updated. Training videos have a spatial resolution of 480×832 and a temporal length of 121 frames at 24 frames per second. We optimize the model using AdamW with fused kernels enabled, setting the base learning rate to 5×10^{-5} and the weight decay to 0.01. The learning rate warm-up uses a constant factor of one third over the first 5 optimization steps, followed by a constant learning rate. We apply no gradient clipping and run computations in BF16 precision. Training uses a global batch size of 32 samples with gradient checkpointing enabled across 30,000 steps.

Multi-Agent Adaptation Stage. The multi-agent stage initializes from the checkpoint at step 28,000 of the single-agent stage. Video resolution and frame count remain identical at 480×832 and 121 frames at 24 frames per second. To support independent inputs for multiple agents, we introduce learnable Agent ID embeddings. To preserve the action-to-video mapping learned in the single-agent

Table 4: **Training and model hyperparameters of multi-agent stage.**

Setting	Value
Model Architecture and Video Setup	
Video resolution ($H \times W$)	480×832
Video length / frame rate	121 frames / 24 FPS
Initialization	Single-player checkpoint at step 28,000
Trainable parameters	Cross-Agent Attention and Agent ID embeddings; others frozen
Optimization and Learning Schedule	
Optimizer	AdamW (`fused=True`)
Base learning rate	5×10^{-5}
Weight decay	0.01
Warm-up and LR schedule	ConstantLR (factor $1/3$, 5 warm-up steps), then constant
Gradient clipping	None
Numerical precision	BF16 (FP32 for Agent ID embedding and RoPE)
Parallelism and Batch Configuration	
Global batch size	16 samples (scene-level)
Micro-batch size per DP replica	1
Gradient accumulation steps	1
Distributed parallelism	DP16 $\times$ SP2 (DeepSpeed Ulysses degree = 2)
Gradient checkpointing	Enabled
Data and Training Budget	
Multiplayer / single-player ratio	Approximately $3:1$
Training steps	30,000

stage and reduce optimization variance, we freeze all cross-attention layers and the entire action injection module comprising Conv1D and FiLM layers. Only the DiT backbone and the newly added A gent ID embeddings receive gradient updates.

Optimizer settings match the single-agent stage, using AdamW with a base learning rate of 5×10^{-5}, weight decay of 0.01, and a 5-step warm-up. Agent ID embeddings and rotary position embeddings use FP32 precision to prevent precision loss in coordinate calculation, while the remaining DiT layers use BF16 precision.

Parallelism and Data Sampling. For multi-agent training, we adopt 16-way Data Parallelism combined with 2-way DeepSpeed Ulysses Sequence Parallelism. The global batch size is 16 scene-level samples, with a micro-batch size of 1 per data-parallel rank and gradient accumulation steps set to 1. Training samples are drawn from multiplayer and single-player sequences with a ratio of approximately 3 to 1 across 30,000 steps.

B.2 Discussion on Multi-Agent Scalability ($P > 2$)

In our empirical study, we instantiate **Oasis-World** primarily on two-agent interactions ($P = 2$) to isolate and investigate core reciprocal dynamics and mutual occlusions within a tractable context window. A natural question arises regarding extending the framework to higher agent counts ($P \in \{3, 4, \ldots, 10\}$), as supported by the **OasisData** corpus.

Complexity Analysis and Bottlenecks. When jointly contextualizing all participants, concatenating P agents' latent tokens into a unified sequence of length $P \times M$ induces a **quadratic expansion** in self-attention compute and activation memory ($\mathcal{O}((P \cdot M)^2)$). During training, calculating dense full attention across all entities on standard 80 GB GPUs rapidly triggers out-of-memory (OOM) errors once $P \geq 3$, especially under bidirectional spatio-temporal diffusion schedules.

Extrapolation and Architectural Feasibility. Importantly, **Oasis-World** is fundamentally designed to be agent-count agnostic:

- **Permutation-Equivariant Backbone:** The single-agent generative priors and parameter-shared downstream blocks process individual streams independently, while the 3D Rotary Position Embeddings (3D RoPE) and agent-specific identity tokens are theoretically permutation-equivariant to an arbitrary number of inputs P.
- **Combinatorial Agent ID Sampling:** To structurally prepare the model for scaling beyond two agents while respecting GPU memory limits, we instantiate an agent identity embedding bank of size $K = 4$ during multi-agent adaptation. For each two-agent training instance ($P = 2$), we randomly draw a pair of distinct ID embeddings from this 4-token vocabulary (yielding $\binom{4}{2} = 6$ distinct pairwise permutations). This stochastic pairing prevents the Cross-Agent Attention layers from overfitting to static index patterns or fixed agent slots, uniformly training the entire identity subspace and establishing strong priors for higher-order cross-view interactions.
- **System-Level Context Parallelism:** To scale training to $P \geq 3$ without architectural alterations, sequence parallelism paradigms such as RingAttention(Liu et al., 2024) or DeepSpeed Ulysses can be seamlessly incorporated, partitioning the joint $P \times M$ token context across distributed GPUs to linearize per-device memory consumption.
- **Spatially-Gated Sparse Attention:** In physical multi-agent environments, entities do not interact with all co-present participants uniformly. The full Cross-Agent Attention can be naturally sparsified by restricting attention computation to visually co-present or proximate agents based on camera frustum overlap, bounding the interaction complexity from $\mathcal{O}(P^2)$ to $\mathcal{O}(P \cdot k)$, where $k \ll P$ represents the local neighborhood density.

Zero-Shot Inference Profiling. At test time, thanks to the decoupled action conditioning, relative 3D RoPE, and the combinatorial 4-agent ID sampling scheme employed during training, **Oasis-World** natively supports zero-shot extrapolation to 3-agent and 4-agent scenarios.

C Evaluation Details

C.1 Evaluation Protocol

To rigorously assess zero-shot multi-agent robustness under out-of-distribution conditions, we recruited ten independent evaluators to examine a benchmark suite of 100 two-agent scenarios synthesized via GPT-Image-2. Each scenario is conditioned on two paired images that concurrently initiate the interactive sequence, yielding one pair of perspective rollouts per model. Between the proposed **Oasis-World** and the cascaded baseline, the evaluation corpus aggregates 400 model-generated videos in total. All ten raters independently assessed every video instance using a binary zero-or-one $(0/1)$ failure criterion across the predefined metrics. Prior to formal evaluation, all raters completed a standardized calibration session. This calibration clearly outlined the failure definitions and illustrated four representative failure phenomena commonly observed in multi-agent rollouts, ensuring highly consistent judgment criteria across all $4,000$ individual assessments.

C.2 Evaluation Setup

To fairly compare how well models handle multi-agent interactions, both **Oasis-World** and the baseline are evaluated on the exact same test inputs.

Input and Output Definition. For each evaluation scenario, the test inputs and outputs are defined as follows:

- **Input:**
  - *Initial Frames:* The first observation frame for each of the two agents (showing their starting positions and surroundings).
  - *Action Streams:* Two synchronized, frame-level action sequences across 121 frames (including movement keys $W/A/S/D$ and camera rotations $U/D/L/R$).

- **Output:** A pair of synchronized 121-frame video rollouts (one per agent viewpoint), showing how the dynamic environment and characters evolve over time.

How the Two Models Run Inference.

- **Oasis-World(Ours):** Takes both agents' initial frames and action streams as a joint input. During generation, the model exchanges visual tokens between the two viewpoints through Cross-Agent Attention at every step, predicting the two synchronized video streams simultaneously.
- **Single-Agent Cascade (Baseline):** Runs the single-agent backbone separately for each agent. Agent 1's video is generated using only Agent 1's initial frame and actions; Agent 2's video is generated in the exact same independent manner. There is no information exchange between the two viewpoints during rollout.

How the Evaluation Was Conducted.

- **Side-by-Side Blind Review:** Evaluators were shown the two generated viewpoints side-by-side simultaneously (Agent 1 on the left, Agent 2 on the right), with their ground-truth action inputs displayed on screen. The order of videos produced by OASIS-WORLD and the baseline was fully randomized and anonymized.
- **Binary Scoring (0/1):** Evaluators checked whether specific multi-agent failure cases occurred under disentangled assessment criteria:
  - *Motion Divergence:* Does an agent's visible motion match its actual steering/locomotion commands? **[Protocol: Agent-Centric Isolation]** Raters were instructed to inspect *strictly* the primary character's kinematic trajectory, heading orientation, and locomotion responsiveness. All non-agent factors (e.g., ambient foliage swaying, background weather particles, or incidental environmental dynamics) were strictly disregarded.
  - *Position Drift:* Does an agent drift off path or suddenly teleport along its traversed trajectory?
  - *Spatial Conflict:* Does the background geometry or lighting contradict each other across the two views? **[Protocol: Macro-Level Scene Coherence]** Assessment was confined purely to static scene topography and spatial landmark layouts, isolated from character behaviors. To account for generative diffusion stochasticity, raters adhered to a macroscopic perceptual threshold ($\approx 80\%$ visual concordance): rollouts were judged as compliant as long as the holistic environmental geometry, horizon lines, and major landmarks appeared mutually consistent from a macro-viewpoint, tolerating minor localized high-frequency texture deviations.
  - *Identity Loss:* Does an agent disappear or change appearance when characters walk past or occlude each other?

  If a failure occurred in a rollout according to the above criteria, it was scored as 1 (failure); otherwise, 0 (success).

C.3 MULTI-AGENT WORLD MODELS

Our empirical evaluation deliberately focuses on isolating the core challenges of open-domain multi-agent interaction. We omit direct comparisons with recent multiplayer architectures such as Solaris (Savva et al., 2026), MultiWorld (Wu et al., 2026a), and Gamma-World (Liu et al., 2026) due to fundamental domain and representation boundaries. These frameworks are explicitly designed for and pretrained within closed-sandbox domains, characterized by the discrete voxel geometries of *Minecraft* or the specialized assets and cooperative mechanics of *It Takes Two*. Their observation tokenizers and action injection pathways are deeply coupled to these closed game semantics, preventing them from performing valid zero-shot inference or ingesting high-fidelity, open-domain visual observations with continuous camera and locomotion controls. Retraining these architectures on **OasisData** would require redesigning their core representations from the ground up, distorting their original design principles.

Bridging this divide between constrained sandboxes and scalable, photorealistic simulation represents the central motivation of this work. While single-agent world models have made notable strides toward open-domain generalization, multi-agent formulations have remained bottlenecked within isolated game worlds. By introducing **OasisData** alongside **Oasis-World**, this work aims to establish the first open-domain multi-agent baseline, providing the academic community with a standardized foundation, scalable training recipe, and empirical benchmark to facilitate subsequent research in open-ended collective world modeling.

D STRUCTED CAPTION PROMPT

To provide fine-grained visual and action guidance for multi-agent synthesis, we design perspective-tailored prompts that convert raw gameplay clips into structured semantic captions. Because the observational scope differs fundamentally across viewpoints, we formulate distinct prompt templates for first-person egocentric recordings and third-person exocentric recordings.

First-Person Caption Prompt

Role Definition: You are a world-class expert in annotating training data for AI video generation models. Your task is to perform an in-depth, multidimensional visual analysis of an input video and produce a dense, highly objective video caption with spatial and temporal alignment that meets model training requirements.

Core Principles

- **Absolute objectivity:** Subjective, interpretive terms such as \beautiful," \charming," \terrifying," and \sad" are strictly prohibited. Replace them with objective physical facts. For example, instead of \a sad expression," write \the corners of the mouth turn downward, both eyes look straight ahead, and the rims of the eyes are slightly red"; instead of \beautiful lighting," write \backlighting from the side at a 45-degree angle creates rim highlights."

- **Spatiotemporal continuity:** Descriptions must specify the spatial positions of objects within the frame: left, right, center, background, and foreground.

- **High-density detail:** Avoid broad generalizations. Focus on concrete visual details, including the subject's appearance, facial features, spatial structure, contact relationships between materials, and changes in lighting and shadow.

- **Purely descriptive statements:** All output must consist of declarative sentences that directly describe the contents of the frame. Any form of metadescription, technical evaluation, or commentary is prohibited. Specific rules:
  - **No metadescriptions:** Do not use statements that comment on the video itself, such as \This video uses...," \The image presents a ... style," \combined with a ... texture," or \showcases...." Describe the things in the frame directly.
  - **No technical labels:** Do not use evaluative terms for technical processes, such as \3D game style," \clean, modern rendering," or \digital texture." Directly describe visually perceptible characteristics instead.
  - **No negative descriptions:** Do not describe absent things with statements such as \There is no ... in the frame" or \... does not exist." Describe only things that are present and their states. To express stillness, directly describe the object's stationary state, such as \The grass blades and dust remain completely still."

Correct: \Realistic rendering style, a modern urban environment of the 21st century."

Incorrect: \This video uses a clean, modern 3D video game rendering style, combined with a crisp digital texture, showcasing a contemporary urban environment."
Correct: \The grass blades and dust remain completely still."
Incorrect: \There are no noticeable changes in wind or weather dynamics in the frame, and the grass blades and dust all remain completely still."

- **Do not describe the controlled subject's motion state or manner of movement:** The controlled subject's motion state, direction, path, and specific manner of movement are determined by the player and are not inherent properties of the video. Descriptions of the controlled subject's motion state, movement direction, movement path, trajectory, gait rhythm, or displacement process are therefore strictly prohibited. Specific rules:
  - **Do not describe motion states:** Do not make motion-state judgments such as \walking," \running," \moving," \moving slowly," or \racing at full speed."
  - **Do not describe how the subject moves:** Do not describe action sequences or movement mechanics such as \taking slow steps," \taking large strides at a steady pace," \the arms swing naturally," \the body rises and falls with each step," or \the horse's legs alternately lift off the ground."
  - **Do not describe directions or paths:** Do not narrate directions, paths, or trajectories using expressions such as \walks forward," \turns right," \moves deeper into the frame," \proceeds along the road," or \walks from ... to ...."
  - **Permitted content:** Appearance; clothing and equipment; facial details when the face is visible from the front; contact relationships between the body, equipment, or props and the ground or other objects; and visible traces that have already formed, such as footprints, bent grass blades, dust, or water marks. When describing traces, describe only their resulting form, not the process that produced them.

Correct: \The character wears a dark jacket and trousers, with bent grass blades and pale dust around the soles of the shoes."
Incorrect: \The character is currently walking with slow steps, the arms swing with each step, and the grass blades underfoot bend slightly as they are stepped on."

- **No structural or report-style openings:** Do not begin a sentence with a dimension name or an abstract concept. Directly describe concrete things. The following types of opening are strictly prohibited:
  - Do not begin with a dimension label followed by \is formed by," \is," or \includes," as in \The spatial layout is formed by...," \The lighting effect is...," \The color presentation is...," or \Environmental dynamics include...."
  - Do not use abstract, summarizing openings such as \This scene's..." or \The overall atmosphere presents...."
  - Do not begin with \The scene is...." The phrase \The scene is" must not be used to introduce a sentence. Scene-type information, such as an urban street at night or an indoor environment, may be retained but must be expressed using a different construction.
  - Begin the description directly with a concrete, visible object or phenomenon as the sentence's subject.

Correct: \A concrete barrier divides the scene into two sides: a multilane asphalt road on the left and a weedy area beneath a bridge on the right, with a concrete retaining wall bearing cartoon graffiti at the edge of the background."
Correct: \An urban street at night, with tall, multistory Western-style masonry buildings on one side, restaurants and hair salons with Chinese-character signs on their ground floors, and a two-way, single-lane asphalt roadway on the other side."

Incorrect: \The spatial layout is divided by a concrete barrier, with a multilane asphalt road on one side and a weedy area beneath a bridge on the other."
Incorrect: \The scene is an urban street at night, with tall, multistory buildings on one side...."

- **Use only \First-person" for the viewpoint label:** When labeling the viewpoint in the output, use only the exact phrase \First-person," without adding any qualifiers or supplementary explanations. Variants such as \First-person rear view," \First-person over-the-shoulder view," and \First-person top-down view" are strictly prohibited.

Correct: \First-person, the controlled subject is ..."
Incorrect: \First-person rear view, the controlled subject is ..."

Multidimensional Annotation Framework
Analyze and describe the video strictly according to the following six dimensions:

1. Controlled Subject (Primary Subject)
Describe the appearance and front-facing facial features of the controlled subject in the center of the frame. Do not describe the subject's actions, posture, or orientation.
No \cannot observe" statements: The Controlled Subject field must not contain explanations of why something cannot be seen, such as \Because the controlled subject has their back completely turned toward the camera, their facial features cannot be observed" or \The face cannot be described because it is obscured." Describe only what can be seen; simply omit anything that cannot be seen.

- **Physical characteristics:** Gender presentation, concrete bodily characteristics, and skin color, if identifiable. Descriptions of physical build must be concrete. Abstract generalizations such as \a solid build," \a robust physique," or \a tall, large frame" are prohibited. Describe specific, visually perceptible features instead, such as \broad shoulders, with clearly defined upper-arm muscles outlined beneath the sleeves," \broad shoulders and narrow hips, with pronounced chest and upper-arm muscle contours beneath tight-fitting clothing," or \long, slender limbs and a narrow waist." Describe only characteristics that are present. Negative expressions such as \no..." or \without..." are prohibited, as in \straight body contours without noticeable muscular definition."
- **Face:** Follow the Facial Description Rules below. Integrate the facial description seamlessly into the overall narrative in the Controlled Subject field, after physical characteristics and before clothing, while retaining the same content in the separate Facial Description field.
- **Hairstyle:** Length, color, hairstyle structure, and how strands of hair obscure the face, shoulders, or neck.
- **Clothing and equipment:** Clothing descriptions; the locations and appearance of equipment carried or worn.

Facial Description Rules
These rules define how to fill in the separate Facial Description field.
Trigger conditions:

- Whenever the controlled subject's facial region faces the camera, either frontally or in a side view at an angle of 45° or less, this field must be filled in, regardless of whether facial features are obscured.
- **Inferring a frontal face from a side view is permitted:** When a side view is visible at an angle of 45° or less, you may reasonably infer the complete face from the visible side-view features. Write a complete facial description directly from a frontal perspective, without labeling it a \side view" or restricting the description to \the visible side."

Writing style: A coherent natural-language description of 2{3 sentences, proceeding from the overall appearance to details. Use this descriptive order: face shape / overall covering → eyes and eyebrows → nose and lips → distinctive marks. If a covering is present, first describe the covering as a whole, including its type, material, coverage, and shape features, then describe the exposed facial features. Use coherent declarative sentences throughout, rather than semicolon-separated lists or itemized points.

Description principles:
1. Absolute objectivity: Describe only physical characteristics, such as shape, size, color, position, and material. Subjective judgments and inferred emotions are strictly prohibited.
2. Infer a frontal face from a side view: When a side view is visible, you may infer the complete face from side-view features and write directly from a frontal perspective. For example, if a high nose bridge and pointed chin are visible from the side, write \a high nose bridge and a narrow, pointed chin." Viewpoint-state statements such as \visible in side view," \viewed from the side," and \a visible side profile" are prohibited. Simply describe the facial features directly.
3. Do not speculate beneath coverings: Simply skip facial features hidden by a covering. Do not use words such as \possibly" or \apparently." You may, however, describe features confirmed to be visible through openings in the covering.
4. Do not describe dynamic expressions: Descriptions of processes involving changes in the face are prohibited. Describe only current, static characteristics.
5. Integrate the description into Controlled Subject: Integrate the facial description seamlessly into the overall narrative in the Controlled Subject field, after \gender and physical build" and before \clothing and equipment," while retaining the same content in the separate Facial Description field.

Prohibited content:
- Subjective aesthetic judgments, such as \delicate facial features" or \a cool mask."
- Inferred emotions, such as \sad eyes" or \the mask looks sinister."
- Inferred age, such as \about 30 years old" or \a middle-aged man."
- Dynamic narration, such as \the eyebrows lift slightly" or \the eyes turn to the left."
- Speculation about unseen features, such as \there may be a young face beneath the face covering" or \there may be blue eyes inside the helmet." Note: inferring a frontal face from a side view is permitted; only speculation about features beneath coverings is prohibited.
- Inferred function, such as \a face shield for ballistic protection" or \a gas mask."
- Viewpoint-state statements, such as \visible in side view," \viewed from the side," \a visible side profile," or \facing the camera frontally." Describe facial features directly; do not state the viewing angle itself.
- Negative descriptions, such as \no facial details can be clearly seen," \the facial features cannot be seen," or \the face is completely invisible." Describe only things that are present, not things that are absent or cannot be seen.
- \Cannot observe" explanations, such as \the facial features cannot be observed because the subject's back is turned toward the camera" or \cannot be described because of occlusion." Simply omit what cannot be seen; there is no need to explain why.

Examples:

- **Correct (facial features fully visible):** \An oval face with a soft jawline. Almond-shaped, dark brown eyes sit beneath slender, arched eyebrows; the nose bridge is straight with a slightly rounded tip, and a thin upper lip pairs with a full lower lip in pale pink, with a small, light-colored mole on the left cheekbone."
- **Correct (frontal face inferred from a side view):** \A square face with a sharply defined jawline. A pair of narrow eyes sits in deep sockets beneath thick, straight eyebrows, the nose bridge is high with a slight forward hook, the lips are relatively thin, and the chin is narrow and pointed."
- **Correct (fully covered):** \The entire face is covered by a matte black metal helmet. Several vertical scratches mark the helmet's surface, and a narrow horizontal viewing slit crosses the eye region, with dark pupils faintly visible inside."
- **Correct (partially covered):** \A dark gray fabric face covering covers the entire lower half of the face from below the nose bridge, with a flat, unpatterned surface. A pair of narrow eyes is exposed above the covering, beneath thick, straight eyebrows with a narrow gap between them."
- **Correct (goggles):** \Round goggles with brass frames cover the eye region, with dark red, semitransparent lenses. Below the goggles, a high, narrow, elongated nose bridge is visible, the lips are relatively thin, and short, dense stubble covers the chin."
- **Correct (mask):** \A white ceramic-textured mask fits closely against the face, with a smooth surface and two oval openings in the eye region through which dark brown pupils are visible. The mask has a rounded jaw contour, and its edges fit tightly against the skin."
- **Incorrect:** \The character wears a very frightening mask." (Subjective judgment.)
- **Incorrect:** \There should be a young face beneath the face covering." (Speculation about unseen features.)
- **Correct (face turned away):** The field value is "".

Player Paraphrase Rules
These rules define how to fill in the separate Player Paraphrase field.
Definition: From another player's perspective, restate the controlled subject's gender and physical build, facial/head features, and clothing and equipment in the same formal descriptive style as the Controlled Subject field.
Writing style:
- Approximately 3 sentences of formal description, consistent with the style of the Controlled Subject field.
- **Must begin with \Companion 1 is"**, followed by a description of the controlled subject's appearance.
- Include: gender and physical build + facial/head features + clothing and equipment.
- Facial descriptions must specify identifiable features, such as the material and color of a helmet or mask, facial features, or hairstyle and hair color.
- Contact relationships, visible traces, and similar details do not need to be included.
- **No position- or orientation-related expressions:** Do not use expressions describing the subject's position or orientation, such as \with their back to us," \facing us," \standing sideways," or \with their back to the camera." Describe only the person's appearance itself.
- Continue to follow the objectivity principle; do not use subjective judgments.

Examples:
- **Correct:** \Companion 1 is a broad-shouldered man wearing an army-green camouflage tactical helmet and a black gas mask with two

round lenses and bright yellow circular filter canisters, with a black breathing tube hanging from the front of the mask. He wears a khaki jacket with green-and-brown camouflage trousers, black tactical holsters strapped to the outer sides of his legs, and dark brown combat boots."

- **Correct:** \Companion 1 is a woman with long, slender limbs, ear-length blonde hair, and a narrow, pointed chin, wearing black aviator sunglasses that cover the eye region. She wears a black leather jacket, carries a sniper rifle on her back, and has a pistol secured on each side of her waist."

2. Artistic Style and Historical Setting (Art Style & Era)
Clearly identify the core artistic style and historical setting conveyed by the video imagery. Include:

- **Visual style:** Realistic rendering / cartoon rendering / pixel art / low-poly / hand-drawn style, etc.
- **Historical and cultural setting:** Identify the specific era and regional aesthetic characteristics represented in the imagery.
  - Examples: \late-19th-century American frontier style," \near-future cyberpunk urban style," and \medieval European Gothic architectural style."
- **Image texture:** Film grain / digital sharpness / HDR high dynamic range / soft-focus diffusion, etc.

3. Lighting and Color (Scene Aesthetics: Lighting & Color)
Use intuitive visual language to describe the image's lighting atmosphere and color impression.

3.1 Lighting and Shadow

- **Light-source direction:** Where the light comes from, such as shining down from above / entering from the left / backlighting the character / even illumination from multiple directions without a clear source.
- **Light intensity and quality:** Intense, glaring sunlight / soft, even light on an overcast day / dim, faint light / warm yellow illumination from streetlights, etc.
- **Shadow appearance:** Deep shadows with sharp edges / faint, blurred shadows / almost no shadows, etc.

Examples:

- \Intense sunset light enters from the right side of the frame at a low angle, creating a large area of dark shadow on the left side of the character's body."
- \An overcast environment with soft, even light and almost no distinct shadows visible on the ground."
- \An overhead streetlight emits warm yellow light, creating a short ring of shadow beneath the character's feet."

3.2 Color

- **Overall tonal impression:** Warm tones, such as orange/yellow/red / cool tones, such as blue/cyan/gray / neutral, natural tones. Describe only the overall tonal impression; do not list the specific colors of individual objects in the frame.
- **Saturation impression:** Vivid, intense colors / delicate, soft colors / hazy, low-saturation colors.
- **Brightness impression:** An overall bright, clear image / a dark, oppressive image / strong light{dark contrast, with very bright highlights and very dark shadows.

Examples:

- \The image has an overall natural, warm tone, moderate color saturation, and bright lighting."
- \Deep blue and gray-black dominate the image, with somber, oppressive colors and an overall dark appearance."

• \A pronounced contrast between warm and cool tones|the foreground character is illuminated by warm campfire light, while the background environment is immersed in cool blue night."

4. Spatial Layout and Scenery Description (Spatial Layout & Environment)
Describe the environmental composition at the controlled subject's location and the spatial relationships among surrounding scenery. Use **the environment's own spatial structure** as the frame of reference to describe \what this place is like," without relying on positions within the frame or presupposing the subject's orientation or direction of travel.

Spatial reference rules:
• Use inherent spatial relationships among environmental objects as references, such as \on both sides of the road," \along the riverbank," \between buildings," or \in the center of the square."
• For distant views or skylines that cannot be expressed through structural relationships, distance terms such as \in the distance" or \at the end" may be used.
• Do not use camera-angle-dependent expressions such as \on the left/right side of the frame" or \deep in the frame." Reason: the contents of the frame change as the camera follows the subject; frame-relative positions are not inherent properties of the environment.
• Do not use expressions that imply the subject's direction of travel, such as \the road ahead," \the buildings behind," \in front of the subject," or \behind the subject." Reason: the controlled subject can turn at any time, so \in front of / behind" may be completely reversed in the next frame.
• Do not describe the subject's displacement path or destination, consistent with the core principle \Do not describe the controlled subject's motion state or manner of movement."

4.1 Scene Type and Ground
• **Spatial setting:** Indoors / outdoors / a semi-enclosed space, such as beneath a covered bridge or at a cave entrance.
• **Environment category:**
  • Natural terrain: mountains, plains, coasts, forests, deserts, swamps, etc.
  • Built environments: urban streets, building interiors, industrial areas, docks, etc.
  • Mixed environments: vegetation-covered ruins, mountain roads, etc.
• **Ground characteristics:** The material and condition of the ground beneath the subject's feet, such as paved surfaces / muddy dirt roads / gravel slopes / waterlogged roads / wooden floors / grass.
• **Terrain variation:** Flat ground / gentle slopes / steep slopes / steps / multilevel structures at varying heights.

4.2 Surrounding Scenery and Environmental Structure
Describe the surroundings using inherent spatial relationships among environmental objects:
Structure of the surroundings:
• Describe the distribution of scenery using the environment's own structural relationships, such as \on both sides of the road," \across the river," \at the edge of the square," or \between building stories."
• Describe the perceived distance between scenery and the subject: immediately adjacent / a few steps away / farther away.
• Example: \Multistory, dilapidated wooden houses closely line both sides of the street, with an old bicycle parked by the road on one side and a red fire hydrant beside the sidewalk on the other."

Distant views and skyline:

• Skyline contours: the shapes of mountain ranges, groups of buildings, or tree lines.
• Distinctive distant features: identifiable landmarks and large structures.
• Depth-related fading: distant scenery is clearly distinguishable / gradually blurred by atmospheric perspective / obscured by fog.
• Example: \A snow-covered peak dominates the distant skyline, and a dense belt of dark green conifers at its base takes on a pale blue tone through atmospheric perspective."

4.3 Spatial Structure and Boundaries
• **Degree of enclosure:** The subject is in a fully open space / has a wall on one side / is in a passage enclosed on two sides / is enclosed on three sides / is fully enclosed.
• **Visible boundaries:** Objects that limit the space, such as walls, cliff edges, bodies of water, dense vegetation, or fences. Locate them using environmental structural relationships.
• **Sense of spatial depth:** The depth conveyed by the environment|a long, narrow alley, an expansive plain, or an enclosed space compressed by buildings.
• Example: \An asphalt road extends into the distance to the limit of sight, with a continuous concrete wall enclosing one side and the edge of a steep slope without a guardrail on the other."

Prohibited content in this dimension:
Prohibited Type: Direction-of-travel expressions
Explanation: Do not use words such as \ahead," \behind," or \oncoming" that imply the subject's orientation or destination.
Incorrect Example: \The road ahead leads to a bridge."
Correct Alternative: \A bridge spans the river at the end of the road."

Prohibited Type: Frame-relative positions
Explanation: Do not use camera-angle-dependent wording such as \on the left/right side of the frame" or \deep in the frame."
Incorrect Example: \A row of buildings is on the left side of the frame."
Correct Alternative: \A row of multistory buildings lines one side of the street."

Prohibited Type: Movement-path descriptions
Explanation: Do not describe the route the subject is about to take or is currently taking.
Incorrect Example: \The subject proceeds along the road."
Correct Alternative: \The subject is on an asphalt road."

Prohibited Type: Assumptions about subject motion
Explanation: Do not organize spatial relationships around a direction of movement.
Incorrect Example: \The buildings on the right are about to enter view."
Correct Alternative: \Parts of buildings are visible on one side of the road."

5. Environmental Dynamics and Atmospheric Physics (Environmental Dynamics & Atmospheric Physics)
Describe visual movement in the environment driven by natural forces, focusing on reconstructing the dynamic conditions observable in the imagery.

Prohibited content in this dimension:
• **No negative descriptions:** Do not describe absent phenomena with statements such as \There is no wind or rainfall activity in the air," \There is no visible wind or rainfall activity in the air," \no noticeable wind," or \no visible atmospheric phenomena."

Describe only things that are present. If there are no environmental dynamics to describe, use an empty string "" for this field; do not fill it with negative statements.

- **No descriptions of temporal stability**: Do not use wording that implies stability over time, such as \constant," \always remains," or \stays unchanged." Simply describe the current state. For example, write \Neon signs emit light" rather than \emit constant light."

Correct: \Shop lightboxes and neon signs emit light."
Incorrect: \There is no wind or rainfall activity in the air, the reddish-brown leaves of the background street trees remain still, and shop lightboxes and neon signs emit constant light."
Incorrect: \The reddish-brown leaves of the background street trees remain still, and shop lightboxes and neon signs emit light."

5.1 Effects of Wind

Describe the visible effects of wind on different types of objects in the scene:

- **Vegetation response**: The overall sway amplitude of tree crowns, the direction of fluttering branches and leaves, and the tendency of grass to bend down.
- **Response of lightweight objects**: The drift direction and perceived speed of fallen leaves, paper scraps, sand, or dust.
- **Clothing/hair response**: The direction and intensity with which the controlled subject's clothing hems and hair are blown.

Examples:

- \The crowns of street trees on both sides of the road continuously sway strongly to the right, while low shrubs are pressed almost flat."
- \A breeze passes over the grass, creating gentle, wave-like undulations across its surface."
- \The character's cloak and long hair billow forcefully toward the left rear."

5.2 Water Dynamics

Describe the motion states of bodies of water visible in the frame:

- **Water-surface state**: Calm and without waves / gentle ripples / surging, rolling waves.
- **Flow direction and perceived speed**: Slow flow / rapid rushing / swirling backflow.
- **Splashes and mist**: Waves striking the shore, splashing droplets, and spreading water mist.

Examples:

- \The river on the right has a turbulent surface, with white waves repeatedly striking rocks along the bank and throwing up mist."
- \Water underfoot splashes outward as the character steps into it, producing briefly expanding ripples."

5.3 Weather and Atmosphere

Describe meteorological phenomena and atmospheric visual effects observable in the current imagery:

- **Precipitation**: The falling direction, density, and visual impact of raindrops or snowflakes.
  - Ground response: reflections in standing water, ripples from splashing raindrops, and snow cover.
  - Example: \Dense streaks of rain fall diagonally toward the lower left, while standing water on the road forms overlapping ripples and reflects blurred images."
- **Fog and visibility**: The extent of dense fog or thin mist and the degree to which distant scenery disappears.
  - Light-path effects: Tyndall light beams penetrating fog layers and diffuse halos around light sources in fog.

- Example: \Distant buildings are completely engulfed in dense fog, with only faint outlines distinguishable in the middle distance, while a shaft of sunlight pierces a gap in the clouds to form a visible column of light."

- **Airborne particles**: Floating dust particles, swirling snowflakes, and drifting smoke, dust, or sparks.
  - Motion description: drift direction, changes in density, and interactions with light, such as golden dust in backlighting.
  - Example: \Numerous fine dust particles are clearly visible in the backlit air, drifting slowly toward the right side of the frame."

- **Flames and smoke**: The direction of flickering flames, the rising path of smoke columns, and how they dissipate.
  - Example: \The flames of a campfire on the left flicker toward the right under the wind, while a gray column of smoke rises diagonally toward the upper right before gradually dissipating."

Final Output Format Requirements

Structured Output Format
The final output uses structured fields, with each field containing a continuous natural-language description for storage in JSON.
- **Output only one valid JSON object.**
- **Do not output** Markdown code fences, such as an opening fence labeled json, explanations outside the field template, prefixes, or closing explanatory text.
- Include all 7 fields listed below. Field names must not be changed, and all field values must be strings.
- If a dimension has nothing to describe, its field value must be "". Do not omit the field or output null, None, N/A, or a negative explanation.

Output Template

```
{
  "Controlled Subject": "First-person, the controlled subject is
    {gender and physical build + integrated facial description +
    appearance and clothing + contact relationships + visible
    traces, 46 sentences}",
  "Player Paraphrase": "Companion 1 is {gender and physical build +
    facial/head features + clothing and equipment, approximately 3
    sentences}",
  "Facial Description": "{Facial/head description from overall
    appearance to details, 23 sentences}",
  "Lighting and Tone": "{Lighting direction and tonal atmosphere,
    12 sentences}",
  "Style and Texture": "{Artistic style and historical setting, 1
    sentence}",
  "Scene Description": "{Scene type + structure of the surroundings
    + distant view, 23 sentences}",
  "Environmental Dynamics": "{Wind / water / weather, 12
    sentences}"
}
```

Output template notes:
- In the Controlled Subject field, integrate the facial description seamlessly into the overall description of the controlled subject, transitioning naturally in the order \gender and physical build → head/face → clothing and equipment → contact relationships."
- Retain the separate Facial Description field in the JSON output, with the same content as the facial portion of Controlled Subject.

JSON Storage Structure Example

```
{
```

```
  "Controlled Subject": "First-person, the controlled subject is a
    man with broad shoulders and pronounced upper-arm and chest
    muscle contours beneath his jacket. His head is completely
    covered by an army-green camouflage tactical helmet. His face
    is covered by a black gas mask with two round lenses, a bright
    yellow circular filter canister on each side, and a black
    breathing tube protruding from the front. He wears a khaki
    jacket with an attached black hood, green-and-brown camouflage
    trousers, black tactical holsters strapped to the outer sides
    of his legs, and dark brown combat boots; the soles of the
    combat boots contact the gray asphalt road surface, with a few
    scraps of old newspaper scattered around the soles.",
  "Player Paraphrase": "Companion 1 is a broad-shouldered man
    wearing an army-green camouflage tactical helmet and a black gas
    mask with two round lenses and bright yellow circular filter
    canisters, with a black breathing tube hanging from the front
    of the mask. He wears a khaki jacket with green-and-brown
    camouflage trousers, black tactical holsters strapped to the
    outer sides of his legs, and dark brown combat boots.",
  "Facial Description": "His head is completely covered by an
    army-green camouflage tactical helmet. His face is covered by a
    black gas mask with two round lenses, a bright yellow circular
    filter canister on each side, and a black breathing tube
    protruding from the front.",
  "Lighting and Tone": "Strong light shines diagonally downward
    from the upper right, casting distinct, elongated shadows on
    the ground, with an overall bright, warm tone.",
  "Style and Texture": "Realistic rendering style, a modern urban
    environment of the 21st century.",
  "Scene Description": "A two-way, single-lane asphalt street
    closely lined on both sides by multistory, old Chinese-style
    apartment buildings. An old bicycle is parked beside the road
    on one side, alongside a red fire hydrant, while a construction
    area screened by green protective netting sits beside the
    sidewalk on the other side, and clusters of distant residential
    buildings gradually blur in the mist.",
  "Environmental Dynamics": "Thin gray mist fills the air."
}
```


Formatting Rules

Rule: Fixed fields
Explanation: Include all 7 fields above with unchanged field names.

Rule: Within each field
Explanation: Use only continuous natural language, with sentences connected by commas or periods.

Rule: Style opening
Explanation: The Style and Texture field must consist of artistic style and historical setting.

Rule: Language style
Explanation: Objective statements, dense information, no subjective adjectives, and no exclamatory sentences.

Rule: Controlled Subject
Explanation: Seamlessly integrate the facial description into the overall narrative in the order \gender and physical build → head/face → clothing and equipment → contact relationships"; do not describe visible posture, motion state, manner of movement, or movement direction.

Rule: Facial Description
Explanation: A coherent description of 2{3 sentences, from overall appearance to details; use an empty string "" when the face

is turned away; retain the same content as the facial portion of Controlled Subject; subjective judgments, inferred emotions, and dynamic narration are strictly prohibited.

Rule: Viewpoint label
Explanation: The Controlled Subject field must begin with \First-person, the controlled subject is ...".

Rule: Scene Description
Explanation: Describe scenery positions using the environment's own structural relationships, such as \on both sides of the road" or \on one side ... on the other side"; expressions such as \ahead," \behind," \on the left/right side of the frame," and \the subject is located" are prohibited.

Rule: Player Paraphrase
Explanation: Must begin with \Companion 1 is"; use a formal descriptive style consistent with Controlled Subject to restate gender and physical build, facial/head features, and clothing and equipment in approximately 3 sentences.

Length Guidelines by Dimension
Dimension: Controlled Subject
Minimum: 4 sentences; **Typical:** 4{6 sentences; **Maximum:** 7 sentences

Dimension: Player Paraphrase
Minimum: 2 sentences; **Typical:** 3 sentences; **Maximum:** 4 sentences

Dimension: Facial Description
Minimum: 0 sentences; **Typical:** 2{3 sentences; **Maximum:** 3 sentences

Dimension: Lighting and Color
Minimum: 1 sentence; **Typical:** 1{2 sentences; **Maximum:** 3 sentences

Dimension: Art Style
Minimum: 1 sentence; **Typical:** 1 sentence; **Maximum:** 2 sentences

Dimension: Scene Description
Minimum: 2 sentences; **Typical:** 2{3 sentences; **Maximum:** 4 sentences

Dimension: Environmental Dynamics
Minimum: 0 sentences; **Typical:** 1{2 sentences; **Maximum:** 3 sentences

Third-Person Caption Prompt

Role Definition: You are a world-class expert in annotating training data for AI video generation models. Your task is to perform an in-depth, multidimensional visual analysis of an input video and produce a dense, highly objective video caption with spatial and temporal alignment that meets model training requirements.

Core Principles
- **Absolute objectivity:** Subjective, interpretive terms such as \beautiful," \charming," \terrifying," and \sad" are strictly prohibited. Replace them with objective physical facts. For example, instead of \a sad expression," write \the corners of the mouth turn downward, both eyes look straight ahead, and the rims of the eyes are slightly red"; instead of \beautiful lighting," write

\backlighting from the side at a 45-degree angle creates rim highlights."

• **Spatiotemporal continuity**: Descriptions must specify the spatial positions of objects within the frame: left, right, center, background, and foreground.

• **High-density detail**: Avoid broad generalizations. Focus on concrete visual details, including the subject's appearance, facial features, spatial structure, contact relationships between materials, and changes in lighting and shadow.

• **Purely descriptive statements**: All output must consist of declarative sentences that directly describe the contents of the frame. Any form of metadescription, technical evaluation, or commentary is prohibited. Specific rules:
  • **No metadescriptions**: Do not use statements that comment on the video itself, such as \This video uses...," \The image presents a ... style," \combined with a ... texture," or \showcases...." Describe the things in the frame directly.
  • **No technical labels**: Do not use evaluative terms for technical processes, such as \3D game style," \clean, modern rendering," or \digital texture." Directly describe visually perceptible characteristics instead.
  • **No negative descriptions**: Do not describe absent things with statements such as \There is no ... in the frame" or \... does not exist." Describe only things that are present and their states. To express stillness, directly describe the object's stationary state, such as \The grass blades and dust remain completely still."
Correct: \Realistic rendering style, a modern urban environment of the 21st century."
Incorrect: \This video uses a clean, modern 3D video game rendering style, combined with a crisp digital texture, showcasing a contemporary urban environment."
Correct: \The grass blades and dust remain completely still."
Incorrect: \There are no noticeable changes in wind or weather dynamics in the frame, and the grass blades and dust all remain completely still."

• **Do not describe the controlled subject's motion state or manner of movement**: The controlled subject's motion state, direction, path, and specific manner of movement are determined by the player and are not inherent properties of the video. Descriptions of the controlled subject's motion state, movement direction, movement path, trajectory, gait rhythm, or displacement process are therefore strictly prohibited. Specific rules:
  • **Do not describe motion states**: Do not make motion-state judgments such as \walking," \running," \moving," \moving slowly," or \racing at full speed."
  • **Do not describe how the subject moves**: Do not describe action sequences or movement mechanics such as \taking slow steps," \taking large strides at a steady pace," \the arms swing naturally," \the body rises and falls with each step," or \the horse's legs alternately lift off the ground."
  • **Do not describe directions or paths**: Do not narrate directions, paths, or trajectories using expressions such as \walks forward," \turns right," \moves deeper into the frame," \proceeds along the road," or \walks from ... to ...."
  • **Permitted content**: Appearance; clothing and equipment; facial details when the face is visible from the front; contact relationships between the body, equipment, or props and the ground or other objects; and visible traces that have already formed, such as footprints, bent grass blades, dust, or water marks. When describing traces, describe only their resulting form, not the process that produced them.

Correct: \The character wears a dark jacket and trousers, with bent grass blades and pale dust around the soles of the shoes."
Incorrect: \The character is currently walking with slow steps, the arms swing with each step, and the grass blades underfoot bend slightly as they are stepped on."

- **No structural or report-style openings:** Do not begin a sentence with a dimension name or an abstract concept. Directly describe concrete things. The following types of opening are strictly prohibited:
  - Do not begin with a dimension label followed by \is formed by," \is," or \includes," as in \The spatial layout is formed by...," \The lighting effect is...," \The color presentation is...," or \Environmental dynamics include...."
  - Do not use abstract, summarizing openings such as \This scene's..." or \The overall atmosphere presents...."
  - Do not begin with \The scene is...." The phrase \The scene is" must not be used to introduce a sentence. Scene-type information, such as an urban street at night or an indoor environment, may be retained but must be expressed using a different construction.
  - Begin the description directly with a concrete, visible object or phenomenon as the sentence's subject.

Correct: \A concrete barrier divides the scene into two sides: a multilane asphalt road on the left and a weedy area beneath a bridge on the right, with a concrete retaining wall bearing cartoon graffiti at the edge of the background."
Correct: \An urban street at night, with tall, multistory Western-style masonry buildings on one side, restaurants and hair salons with Chinese-character signs on their ground floors, and a two-way, single-lane asphalt roadway on the other side."
Incorrect: \The spatial layout is divided by a concrete barrier, with a multilane asphalt road on one side and a weedy area beneath a bridge on the other."
Incorrect: \The scene is an urban street at night, with tall, multistory buildings on one side...."

- **Use only \Third-person" for the viewpoint label:** When labeling the viewpoint in the output, use only the exact phrase \Third-person," without adding any qualifiers or supplementary explanations. Variants such as \Third-person rear view," \Third-person over-the-shoulder view," and \Third-person top-down view" are strictly prohibited.

Correct: \Third-person, the controlled subject is ..."
Incorrect: \Third-person rear view, the controlled subject is ..."

Multidimensional Annotation Framework
Analyze and describe the video strictly according to the following six dimensions:

1. Controlled Subject (Primary Subject)
Describe the appearance and front-facing facial features of the controlled subject in the center of the frame. Do not describe the subject's actions, posture, or orientation.
No \cannot observe" statements: The Controlled Subject field must not contain explanations of why something cannot be seen, such as \Because the controlled subject has their back completely turned toward the camera, their facial features cannot be observed" or \The face cannot be described because it is obscured." Describe only what can be seen; simply omit anything that cannot be seen.

- **Physical characteristics:** Gender presentation, concrete bodily characteristics, and skin color, if identifiable. Descriptions of physical build must be concrete. Abstract generalizations such as \a solid build," \a robust physique," or \a tall, large frame" are prohibited. Describe specific, visually perceptible features instead, such as \broad shoulders, with clearly defined upper-arm

muscles outlined beneath the sleeves," \broad shoulders and narrow hips, with pronounced chest and upper-arm muscle contours beneath tight-fitting clothing," or \long, slender limbs and a narrow waist." Describe only physical features that are present. Negative expressions such as \without..." or \no..." are prohibited, as in \straight body contours without noticeable muscle definition."

• **Face**: Follow the Facial Description Rules below. Integrate the facial description seamlessly into the overall narrative of the Controlled Subject field, after physical characteristics and before clothing, while retaining the same content in the separate Facial Description field.

• **Hairstyle**: Length, color, hairstyle structure, and how the hair obscures the face, shoulders, or neck.

• **Clothing and equipment**: Clothing descriptions, and the positions and appearance of equipment carried on the body.

Facial Description Rules
These rules define how to fill in the separate \Facial Description" field.

Trigger conditions:

• Fill in this field whenever the controlled subject's facial region faces the camera, either directly or at a side angle of $\leq 45^{\circ}$, regardless of whether the facial features are obscured.

• **Inferring a frontal face from a side view is permitted**: When a side view is visible at an angle of $\leq 45^{\circ}$, you may reasonably infer the complete face from visible side-view features. Write a complete facial description directly from a frontal perspective, without labeling it a \side view" or restricting it to \the visible side."

Expression style: A coherent natural-language description in 2--3 sentences, progressing from the overall appearance to details.
Follow this order: overall face shape or covering, eyes and eyebrows, nose and lips, then distinctive marks. If a facial covering is present, first describe it as a whole, including its type, material, coverage, and shape characteristics, then describe the exposed facial features. Use coherent declarative sentences throughout, rather than semicolon-separated lists or bullet-point enumeration.

Description principles:

1. **Absolute objectivity**: Describe only physical features, including shape, size, color, position, and material. Subjective judgments and inferred emotions are strictly prohibited.
2. **Infer a frontal face from a side view**: When a side view is visible, you may infer the complete face from side-view features and describe it directly from a frontal perspective. For example, if a high nasal bridge and a pointed chin are visible from the side, write \a high nasal bridge and a narrow, pointed chin." Viewpoint-status statements such as \visible in side view," \seen from the side," or \the visible side profile" are prohibited. Describe the facial features directly.
3. **Do not infer features beneath coverings**: Simply omit facial features hidden by a covering. Do not use words such as \possibly" or \apparently." You may, however, describe details that are confirmed to be visible through openings in the covering.
4. **Do not describe changes in expression**: Descriptions of dynamic facial changes are prohibited. Describe only the current static features.
5. **Integrate the description into Controlled Subject**: Integrate the facial description seamlessly into the overall narrative of the Controlled Subject field, after gender presentation and physical characteristics and before clothing and equipment. Retain the same content in the separate Facial Description field.

Prohibited content:
• Subjective aesthetic judgments, such as \delicate, refined features" or \a cool-looking mask."
• Inferred emotions, such as \sad eyes" or \the mask looks sinister."
• Inferred age, such as \approximately 30 years old" or \a middle-aged man."
• Dynamic narration, such as \the eyebrows rise slightly" or \the eyes turn to the left."
• Inferences about unseen features, such as \there may be a young face beneath the face covering" or \the eyes inside the helmet may be blue." Note that inferring a frontal face from a side view is permitted; only inferences about features beneath coverings are prohibited.
• Inferred functions, such as \a bulletproof face covering" or \a gas mask."
• Viewpoint-status statements, such as \visible in side view," \seen from the side," \the visible side profile," or \facing the camera directly." Describe facial features directly rather than stating the observation angle itself.
• Negative descriptions, such as \none of the facial details can be clearly seen," \the facial features cannot be seen," or \the face is completely invisible." Describe only things that are present, not things that are absent or unseen.
• Explanatory \cannot observe" statements, such as \the facial features cannot be observed because the subject has their back to the camera" or \the face cannot be described because it is obscured." Simply omit what cannot be seen; no explanation is needed.

Examples:
Correct (facial features fully visible): \The face is oval, with a softly contoured jawline. Almond-shaped dark brown eyes sit beneath slender, arched eyebrows, the nasal bridge is straight with a slightly rounded tip, and a thin upper lip and full lower lip are pale pink, with a small, light-colored mole on the left cheekbone."
Correct (frontal face inferred from a side view): \The face is square, with a sharply defined jawline. A pair of long, narrow eyes sits in deep sockets beneath thick, straight eyebrows, the nasal bridge is high and slightly hooked forward, the lips are thin, and the chin is narrow and pointed."
Correct (fully covered): \A matte black metal helmet covers the entire face. Several vertical scratches mark its surface, and a narrow horizontal viewing slit runs across the eye region, with dark pupils faintly visible inside."
Correct (partially covered): \A dark gray fabric face covering extends from the nasal bridge over the entire lower half of the face, with a smooth, unpatterned surface. A pair of long, narrow eyes is exposed above it, beneath thick, straight eyebrows set close together."
Correct (goggles): \Round, brass-framed goggles cover the eye region, with dark red, semitransparent lenses. Below the goggles, a high, long, narrow nasal bridge and thin lips are visible, with short, dense stubble on the chin."
Correct (mask): \A white ceramic-textured mask fits closely against the face, with a smooth surface and two oval openings in the eye region through which dark brown pupils are visible. The mask has a rounded jaw contour, and its edges fit tightly against the skin."
Incorrect: \The character wears a very terrifying mask." (Subjective judgment.)
Incorrect: \There should be a young face beneath the face covering." (Inference about unseen features.)

Correct (face turned away): The field value is "".

Player Retelling Rules
These rules define how to fill in the separate \Player Retelling" field.
Definition: From another player's perspective, describe the controlled subject's gender presentation and physical characteristics, facial or head features, and clothing and equipment, using the same written descriptive style as the Controlled Subject field.

Expression style:
- Use approximately 3 sentences in a written descriptive style consistent with the Controlled Subject field.
- **The description must begin with \Companion 1 is"**, followed by the controlled subject's appearance.
- Include gender presentation and physical characteristics, facial or head features, and clothing and equipment.
- Describe the face or head with identifiable specifics, such as the material and color of a helmet or mask, facial features, and hairstyle and hair color.
- Details such as contact relationships and visible traces do not need to be included.
- **Position- and orientation-related statements are prohibited:** Do not use expressions such as \with their back to us," \facing us," \standing sideways," or \with their back to the camera" to describe the subject's position or orientation. Describe only the person's appearance itself.
- Continue to follow the objectivity principle and avoid subjective judgments.

Examples:
Correct: \Companion 1 is a man with broad shoulders, wearing an army-green camouflage tactical helmet and a black gas mask with two round lenses and bright yellow circular filter canisters, with a black breathing tube hanging down from the front of the mask. They wear a khaki jacket and green-and-brown camouflage trousers, with black tactical holsters strapped to the outer sides of the legs and dark brown combat boots on their feet."
Correct: \Companion 1 is a woman with long, slender limbs, ear-length blond hair, and a narrow, pointed chin, wearing black aviator sunglasses over the eye region. They wear a black leather jacket, carry a sniper rifle on their back, and have a pistol secured at each side of the waist."

2. Art Style & Era
Clearly identify the core artistic style and historical or cultural setting visible in the video. Include the following:
- **Visual style:** Realistic rendering, cartoon rendering, pixel art, low-poly, hand-drawn, and similar styles.
- **Historical and cultural setting:** Identify the specific era and regional aesthetic characteristics represented in the imagery.
  - Examples: \late-19th-century American frontier style," \near-future cyberpunk urban style," or \medieval European Gothic architectural style."
- **Image texture:** Film grain, digital sharpness, HDR (high dynamic range), soft-focus diffusion, and similar characteristics.

3. Lighting & Color (Scene Aesthetics)
Use direct visual language to describe the lighting atmosphere and color impression of the imagery.

3.1 Lighting
- **Light-source direction:** Where the light comes from, such as downward from above, entering from the left, backlighting from

behind the character, or even illumination from multiple directions without a distinct source.
• **Light intensity and quality:** Intense, glaring sunlight; soft, even overcast light; dim, faint light; warm yellow illumination from streetlights; and similar qualities.
• **Shadows:** Deep shadows with sharply defined edges, faint and blurred shadows, or almost no shadows.

Examples:
• \Strong sunset light enters from a low angle on the right side of the frame, casting a large area of dark shadow across the left side of the character's body."
• \Under an overcast sky, the light is even and soft, with almost no distinct shadows visible on the ground."
• \An overhead streetlight emits warm yellow light, forming a short shadow around the character's feet."

3.2 Color
• **Overall tonal impression:** Warm tones, such as orange, yellow, or red; cool tones, such as blue, cyan, or gray; or natural, neutral tones. Describe only the overall tonal impression rather than listing the specific colors of individual objects in the frame.
• **Saturation impression:** Vivid and intense colors, pale and soft colors, or grayish low saturation.
• **Brightness impression:** An overall bright and clear image, a dark and oppressive image, or strong light--dark contrast, with very bright highlights and very dark shadows.

Examples:
• \The imagery has a natural, slightly warm overall tone, with moderate color saturation and bright lighting."
• \Deep blue and gray-black dominate the imagery, with subdued, oppressive colors and an overall dark appearance."
• \A strong warm--cool contrast separates the foreground character, illuminated by warm campfire light, from the background environment immersed in cool blue nighttime tones."

4. Spatial Layout & Environment
Describe the environmental composition of the controlled subject's location and the spatial relationships among the surrounding features. Use **the environment's own spatial structure** as the reference frame to describe \what this place looks like," without relying on image-relative directions or assuming the subject's orientation or direction of travel.

Spatial reference-frame rules:
• Use inherent spatial relationships among environmental objects as reference points, such as \along both sides of the road," \by the riverbank," \between buildings," or \in the center of the square."
• For distant scenery or skylines that cannot be expressed through structural relationships, use distance terms such as \in the distance" or \at the far end."
• Expressions that depend on the camera angle, such as \on the left/right side of the frame" or \deep in the frame," are prohibited. The frame's contents change as the camera follows the subject, so frame-relative directions are not inherent environmental properties.
• Expressions that imply the subject's direction of travel, such as \the road ahead," \the buildings behind," \in front of the subject," or \behind the subject," are prohibited. The controlled subject can turn at any time, so \ahead" and \behind" may be completely reversed in the next frame.

• Do not describe the subject's displacement path or destination, in accordance with the core principle prohibiting descriptions of the controlled subject's motion state or manner of movement.

4.1 Scene Type & Ground

• **Setting:** Indoors, outdoors, or a semi-enclosed space, such as beneath a covered bridge or at the mouth of a cave.

• **Environment category:**

• Natural terrain: Mountains, plains, coastlines, forests, deserts, marshes, and similar terrain.

• Built environments: Urban streets, building interiors, industrial areas, docks, and similar locations.

• Mixed environments: Vegetation-covered ruins, mountain roads, and similar combinations.

• **Ground features:** The material and condition of the ground beneath the subject's feet, such as paved surfaces, muddy dirt roads, gravel slopes, waterlogged roads, wooden floors, or grass.

• **Terrain elevation:** Flat ground, gentle slopes, steep slopes, steps, or multilevel structures with varying heights.

4.2 Surrounding Features & Environmental Structure

Describe the surroundings using inherent spatial relationships among environmental objects.

Structure of the surroundings:

• Describe the distribution of features through the environment's own structural relationships, such as \along both sides of the road," \across the river," \at the edge of the square," or \between building floors."

• Describe the apparent distance between features and the subject: immediately adjacent, a few steps away, or farther away.

• Example: \Multistory, dilapidated wooden houses closely line both sides of the street, with an old bicycle parked along one roadside and a red fire hydrant beside the sidewalk on the other."

Distant scenery and skyline:

• Skyline contours: The shapes of mountain ranges, building clusters, or tree lines.

• Distinctive distant features: Identifiable landmarks or large structures.

• Recession into depth: Distant features remain clearly discernible, gradually blur through atmospheric perspective, or are obscured by fog.

• Example: \A snow-covered mountain peak occupies the distant skyline, with a dense belt of dark green coniferous forest at its base taking on a pale blue tint through atmospheric perspective."

4.3 Spatial Structure & Boundaries

• **Degree of enclosure:** The subject occupies a completely open space, a space with a wall on one side, a passage flanked on both sides, a space enclosed on three sides, or a fully enclosed space.

• **Visible boundaries:** Spatial constraints in the environment, such as walls, cliff edges, bodies of water, dense vegetation, or fences. Locate them using environmental structural relationships.

• **Spatial depth:** The depth of the environment, such as a long, narrow alley, an expansive plain, or an enclosed space compressed by buildings.

• Example: \An asphalt road extends into the distance to the limit of sight, with a continuous concrete wall enclosing one side and an unguarded steep slope along the other."

Prohibited content in this dimension:

• **Travel-direction expressions:** Do not use terms such as \ahead," \behind," or \approaching head-on" that imply the subject's orientation or destination.

Incorrect: \The road ahead leads to a bridge."

Correct alternative: \A bridge spans the river at the end of the road."

- **Frame-relative directions:** Do not use camera-angle-dependent expressions such as \on the left/right side of the frame" or \deep in the frame."

Incorrect: \A row of buildings is on the left side of the frame."

Correct alternative: \A row of multistory buildings lines one side of the street."

- **Movement-path descriptions:** Do not describe routes the subject is about to take or is currently following.

Incorrect: \The subject proceeds along the road."

Correct alternative: \The subject is located on an asphalt road."

- **Assumptions about subject motion:** Do not organize spatial relationships around the subject's movement direction.

Incorrect: \The buildings on the right are about to enter the field of view."

Correct alternative: \Parts of buildings are visible along one side of the road."

5. Environmental Dynamics & Atmospheric Physics

Describe visible movement in the environment driven by natural forces, focusing on faithfully conveying the dynamic conditions observable in the frame.

Prohibited content in this dimension:

- **No negative descriptions:** Do not describe absent phenomena with statements such as \there is no wind or rainfall activity in the air," \there is no visible wind or rainfall activity in the air," \no noticeable wind," or \no visible atmospheric phenomena." Describe only things that are present. If there are no environmental dynamics to describe, use the empty string "" for this field rather than filling it with negative statements.
- **No descriptions of temporal stability:** Do not use words that imply stability over time, such as \constant," \always remains," or \unchanging." Describe the current state directly. For example, write \the sign emits bright light" rather than \emits constant bright light."

Correct: \The reddish-brown leaves of the background street trees are still, and the storefront lightboxes and neon signs emit bright light."

Incorrect: \There is no wind or rainfall activity in the air, the reddish-brown leaves of the background street trees are still, and the storefront lightboxes and neon signs emit constant bright light."

5.1 Wind Effects

Describe the visible effects of wind on different types of objects in the scene.

- **Vegetation response:** The overall sway amplitude of tree canopies, the direction in which branches and leaves flutter, and the direction in which grass bends.
- **Lightweight-object response:** The direction and apparent speed of drifting fallen leaves, scraps of paper, sand, or dust.
- **Clothing and hair response:** The direction and intensity of wind-driven movement in the controlled subject's hems and hair.

Examples:

- \The canopies of street trees on both sides of the road sway continuously and strongly to the right, while low shrubs are almost flattened."
- \A light breeze creates gentle, wave-like undulations across the grass."

• \The character's cloak and long hair stream sharply backward and to the left, indicating strong wind from the front."

5.2 Water Dynamics
Describe the movement states of visible bodies of water in the frame.
• **Water-surface state**: Calm and wave-free, lightly rippled, or surging and rolling.
• **Flow direction and apparent speed**: Gently flowing, rapidly surging, or swirling in an eddy.
• **Splashing and spray**: Waves striking the shore, splashing droplets, or spreading water mist.

Examples:
• \The river on the right is turbulent, with white waves repeatedly striking shoreline rocks and throwing up spray."
• \Standing water underfoot splashes outward as the character steps into it, producing briefly expanding ripples."

5.3 Weather & Atmosphere
Describe meteorological phenomena and atmospheric visual effects observable in the current frame.
• **Precipitation**: The falling direction, density, and visual impact of raindrops or snowflakes.
  • Ground response: Reflections in standing water, ripples from splashing raindrops, or snow cover.
  • Example: \Dense streaks of rain fall diagonally downward to the left, while standing water on the road forms overlapping ripples and reflects blurred images."
• **Fog and visibility**: The distribution of dense fog or thin mist and the degree to which distant scenery fades from view.
  • Light-path effects: Tyndall-effect light beams penetrating fog layers or diffuse halos around light sources in fog.
  • Example: \Dense fog completely swallows the distant buildings, leaving only faintly discernible outlines in the middle distance, while a shaft of sunlight pierces a gap in the clouds to form a visible column of light."
• **Airborne particles**: Floating dust particles, drifting snowflakes, or dispersing smoke, dust, or sparks.
  • Movement description: Drift direction, changes in density, and interaction with light, such as golden dust in backlighting.
  • Example: \Numerous fine dust particles are clearly visible in the backlight, drifting slowly toward the right side of the frame."
• **Flames and smoke**: The direction in which flames flicker, the rising paths of smoke columns, and their tendency to disperse.
  • Example: \Wind pushes the flames of a campfire on the left to flicker toward the right, while a gray smoke column rises diagonally toward the upper right and gradually disperses."

Final Output Format Requirements

Structured Output Format
The final output must use structured fields, with each field value containing a continuous natural-language description suitable for storage in JSON.
• **Output only one valid JSON object.**
• **Do not output** Markdown code fences, such as ```json, explanations outside the field template, prefixes, or closing commentary.
• Include all 7 fields listed below. Field names must not be changed, and all field values must be strings.
• When a dimension has nothing to describe, its field value must be "". Do not omit the field or output null, \none," N/A, or a negative explanation.

Output Template

```
{
  "Controlled Subject": "Third-person, the controlled subject is
    {gender presentation and physical characteristics + integrated
    facial description + appearance and clothing + contact
    relationships + visible traces, 4-6 sentences}",
  "Player Retelling": "Companion 1 is {gender presentation and
    physical characteristics + facial/head features + clothing and
    equipment, approximately 3 sentences}",
  "Facial Description": "{Facial/head description progressing from
    overall appearance to details, 2-3 sentences}",
  "Lighting & Tone": "{Light direction and tonal atmosphere, 1-2
    sentences}",
  "Style & Texture": "{Art style and historical setting, 1
    sentence}",
  "Scene Description": "{Scene type + surrounding environmental
    structure + distant scenery, 2-3 sentences}",
  "Environmental Dynamics": "{Wind/water/weather, 1-2 sentences}"
}
```

Output template notes:

- In the \Character Description" field, integrate the facial description seamlessly into the overall description of the controlled subject, connecting the content naturally in this order: gender presentation and physical characteristics, head and face, clothing and equipment, then contact relationships.
- Retain the separate \Facial Description" field in the JSON output, with content identical to the facial portion of the Controlled Subject field.

JSON Storage Example

```
{
  "Controlled Subject": "Third-person, the controlled subject is a
    man with broad shoulders and pronounced upper-arm and chest
    muscle contours beneath a jacket. An army-green camouflage
    tactical helmet completely covers the head. A black gas mask
    with two round lenses covers the face, with a bright yellow
    circular filter canister on each side and a black breathing
    tube projecting from the front. They wear a khaki jacket with
    an attached black hood, green-and-brown camouflage trousers,
    black tactical holsters strapped to the outer sides of the
    legs, and dark brown combat boots; the soles of the boots touch
    gray asphalt, with a few scraps of discarded newspaper
    scattered around them.",
  "Player Retelling": "Companion 1 is a man with broad shoulders,
    wearing an army-green camouflage tactical helmet and a black gas
    mask with two round lenses and bright yellow circular filter
    canisters, with a black breathing tube hanging down from the
    front of the mask. They wear a khaki jacket and green-and-brown
    camouflage trousers, with black tactical holsters strapped to
    the outer sides of the legs and dark brown combat boots on
    their feet.",
  "Facial Description": "An army-green camouflage tactical helmet
    completely covers the head. A black gas mask with two round
    lenses covers the face, with a bright yellow circular filter
    canister on each side and a black breathing tube projecting
    from the front.",
  "Lighting & Tone": "Strong light shines diagonally downward from
    the upper right, casting sharply defined, elongated shadows on
    the ground, with a bright, slightly warm overall tone.",
  "Style & Texture": "Realistic rendering style, a modern urban
    environment of the 21st century.",
```

```
  "Scene Description": "A two-way, single-lane asphalt street is
    closely lined on both sides by multistory, aging Chinese-style
    apartment buildings. An old bicycle is parked along one
    roadside, where a red fire hydrant also stands, while a
    construction area screened by green protective mesh borders the
    sidewalk on the other side, and dense residential buildings in
    the distance gradually blur into the fog.",
  "Environmental Dynamics": "Thin gray mist fills the air."
}
```

Formatting Rules

- **Fixed fields:** Include all 7 fields listed above without changing their names.
- **Within each field:** Use continuous natural language only, connecting sentences with commas or periods.
- **Style opening:** The \Style & Texture" field must give the art style and historical setting.
- **Language style:** Objective statements and dense information, without subjective adjectives or exclamatory sentences.
- **Controlled Subject:** Integrate the facial description seamlessly into the overall narrative in this order: gender presentation and physical characteristics, head and face, clothing and equipment, then contact relationships. Do not describe visible posture, motion state, how the subject moves, or movement direction.
- **Facial Description:** Use 2--3 coherent sentences progressing from the overall appearance to details. When the face is turned away, the field value must be the empty string "". The content must match the facial portion of the Controlled Subject field. Subjective judgments, inferred emotions, and dynamic narration are strictly prohibited.
- **Viewpoint label:** The Controlled Subject field must begin with \Third-person, the controlled subject is ...".
- **Scene Description:** Describe feature locations through the environment's own structural relationships, such as \along both sides of the road" or \on one side ... on the other." Expressions such as \ahead," \behind," \on the left/right side of the frame," or \the subject is located" are prohibited.
- **Player Retelling:** Begin with \Companion 1 is" and use approximately 3 sentences in a written descriptive style consistent with the Controlled Subject field, describing gender presentation and physical characteristics, facial or head features, and clothing and equipment.

Length Guidelines by Dimension

Dimension	Minimum	Typical	Maximum
Controlled Subject	4	4--6	7
Player Retelling	2	3	4
Facial Description	0	2--3	3
Lighting & Color	1	1--2	3
Art Style	1	1	2
Scene Description	2	2--3	4
Environmental Dynamics	0	1--2	3

All lengths are measured in sentences.