SPORTCRAFT: AGENTIC 4D RECONSTRUCTION OF SPORTS SCENES FROM MONOCULAR VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular sports reconstruction must align venue geometry, human motion and equipment interactions despite ambiguous scale, rapid motion and changing contacts. Corrections tailored to one action or view can also fail when reused across videos. We propose SportCraft, an agentic framework for editable 4D sports reconstruction organized around two coupled feedback loops, with Codex serving as the supervisory agent for scene diagnosis, code proposals and update review. Agent-directed model fusion registers complementary pretrained predictions in a shared scene with consistent coordinates and timing. Knowledge-guided scene refinement forms the inner loop, using rendered feedback, venue references and phase-specific sports constraints to refine cameras, human motion, object trajectories and contacts. Visual, geometric and contact checks govern the acceptance of scene updates. Agent-driven tool evolution forms the outer loop: the agent records repair experience, generates and revises parameterized geometric functions with explicit applicability conditions, and uses replay and regression checks to review candidate revisions before reuse. Pretrained model weights remain fixed.

1 INTRODUCTION

Reconstructing editable dynamic scenes from video through code offers a bridge between visual observations and computational models of the physical world. Cameras, scene geometry, human articulation and object motion become explicit variables that programs can inspect, replay and modify, providing a basis for motion analysis, interactive editing and downstream simulation. Recent inverse-graphics agents move toward this goal by translating images into executable scene programs. VIGA (Yin et al., 2026) interleaves code generation, rendering and visual inspection to iteratively reconstruct and edit 3D scenes. SEIG (He et al., 2026) decomposes reconstruction into geometry, materials, composition and lighting stages, each guided by rendered feedback. These approaches recover editable scene structure and appearance and also enable downstream physical simulation. However, recovering the motion actually observed in a video requires temporal evidence to constrain trajectories, articulated motion and changing contacts.

A complementary line of research focuses on recovering time-varying geometry, motion and interactions from monocular video. DreamScene4D (Chu et al., 2024) lifts monocular videos into dynamic multi-object Gaussian scenes by separating camera motion, object deformation and object-to-world motion. OVOW (Chen et al., 2026) constructs instance-level mesh scenes, estimates metric scale and pose trajectories through iterative rendering and matching, and enforces ground contact and inter-object support for downstream simulation. RHINO (Xue et al., 2026) jointly reconstructs a human, a novel manipulated object and the static scene in a shared world frame, using compositional neural fields and differentiable contact priors. HAT-4D (Li et al., 2026) combines multiple agents, interaction knowledge and human feedback to reconstruct object geometry and temporal interactions, including under occlusion. Together, these approaches demonstrate how geometric alignment, physical constraints and agent-guided feedback support structured dynamic reconstruction.

Despite these advances, sports reconstruction faces three coupled challenges. *Geometric inconsistency.* Camera motion and ambiguous venue markings make it difficult to align camera, human and object estimates in a shared metric frame, even with known venue dimensions. *Temporal interaction inconsistency.* Rapid motion, small fast-moving objects and brief contacts can cause independently estimated trajectories to disagree on when and where interactions occur. Maintaining sustained

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



Figure 1: **From monocular videos to editable 4D sports scenes.** badminton and soccer reconstructions combine human motion, equipment and venue geometry. Left: auxiliary scene views. Right: six source frames paired with independently centered athlete views; colors encode time.

contacts while resolving impacts, free flight and bounces therefore requires phase-specific constraints. *Limited generalization of repair procedures.* Corrections developed for specific actions or views may fail when their assumptions no longer hold, such as a grounded-motion repair applied to an airborne serve. Generalization requires explicit applicability conditions and cross-video validation to turn action-specific corrections into reusable tools.

To address these challenges, we propose SportCraft, an agentic framework for editable 4D sports reconstruction with *two coupled feedback loops*. Our reconstructed sports scenes integrate human motion, equipment and venue geometry (Figure 1). Codex serves as the supervisory agent: it diagnoses discrepancies, proposes code changes and reviews candidate updates, while model adapters and numerical executors provide predictions and solve geometric subproblems. We leverage pretrained models for human motion (Shen et al., 2024), scene geometry (Wang et al., 2025) and image localization (Wang et al., 2026b) to initialize a shared scene with aligned coordinates and timestamps. Within each video, the *inner loop* uses rendered feedback to diagnose discrepancies and refines camera pose and scale when sufficiently constrained by identified venue references. With the camera fixed, it refines human motion, ball trajectories and equipment contact using phase-specific sports constraints. Updates are retained only after passing visual, geometric and contact checks; otherwise, they are rolled back. Across development videos, the *outer loop* distills repair experience into parameterized functions with explicit applicability conditions and revises their checks or solvers to accommodate new cases. Revisions must pass replay checks on new videos and regression checks on earlier ones before entering the tool library, enabling subsequent inner loops to reuse validated procedures. Pretrained model weights remain fixed, and memory and tools are frozen for held-out evaluation. Our contributions are:

- **Agent-directed model fusion.** We formulate an editable 4D sports representation that integrates human motion, scene geometry and object localization through shared coordinates and timing, using visually supported court geometry and known-size objects to constrain camera pose and scale.
- **Knowledge-guided scene refinement.** We propose a visual feedback procedure that selects sports constraints for flight, impact and contact phases and jointly refines cameras, ball trajectories, human poses and equipment contact, with coupled geometric checks governing scene updates.
- **Agent-driven tool evolution.** We integrate Codex into the feedback loop to record repair experience and applicability conditions, generate and revise parameterized geometric functions, and review candidate tool updates using replay and regression checks. Pretrained model weights remain fixed throughout this process.

2 Related Work

2.1 Programmatic Inverse Graphics

Programmatic inverse graphics represents scenes as executable graphics programs, allowing visual observations to guide changes to geometry, appearance and layout. BlenderAlchemy (Huang et al., 2024) uses visual feedback to propose and evaluate graphics edits, while BlenderGym (Gu et al., 2025) benchmarks code-based editing and visual verification. Recent reconstruction systems extend this process through staged reasoning and memory. VIGA (Yin et al., 2026) interleaves code execution, rendering and inspection using multimodal memory and reusable skills. SEIG (He et al., 2026) decomposes reconstruction into geometry, materials, composition and lighting stages, while Code-as-Room (Yang et al., 2026) uses staged execution and cross-stage memory for room reconstruction. For dynamic scenes, HAT-4D (Li et al., 2026) incorporates interaction knowledge, cached states and human feedback to recover temporal interactions. These methods establish visual feedback as a mechanism for updating editable scenes. Our framework builds on this direction by coupling sports scene refinement with the revision of reusable geometric repair programs.

2.2 Monocular Motion and Interaction Reconstruction

Monocular reconstruction seeks to recover camera motion, articulated people and their interactions in a common 3D frame. SLAHMR (Ye et al., 2023) jointly estimates camera and human motion, while GVHMR (Shen et al., 2024) recovers world-grounded motion through gravity-view coordinates. Extending reconstruction to human–object interactions, RHINO (Xue et al., 2026) models a human, a manipulated object and the static scene with contact and interpenetration constraints. Sports provide additional geometric and motion priors. TVCalib (Theiner & Ewerth, 2023) and PnLCalib (Gutiérrez-Pérez & Agudo, 2024) use known field geometry for camera calibration, while MonoTrack (Liu & Wang, 2022) combines court geometry, player poses and shuttle dynamics for badminton trajectory recovery. TT3D (Gossard et al., 2025) integrates table calibration, player reconstruction and ball trajectory and spin estimation; TT4D (Rahmanian et al., 2026) further reconstructs racket states at impact. These works provide both component estimators and integrated sports pipelines. Building on their geometric and physical constraints, we focus on selecting constraints for changing interaction phases and checking their effects on the coupled camera, human and object estimates.

2.3 Experience-Driven Skill Acquisition

Experience-driven agents retain feedback and executable procedures so that later tasks can reuse earlier reasoning and actions. Reflexion (Shinn et al., 2023) stores verbal reflections to guide subsequent attempts. Beyond textual memory, Voyager (Wang et al., 2024) accumulates executable skills through interaction and feedback, while CRAFT (Yuan et al., 2024) abstracts task solutions into reusable functions and retrieves them for new problems. In visual reasoning, VADAR (Marsili et al., 2025) generates task-specific APIs, and Transductive Visual Programming (Wu et al., 2025) organizes solved examples and verified tools into complementary libraries. ASPIRE (Lu et al., 2026) uses multimodal execution traces to diagnose failures and revise robotic skills. Together, these approaches connect stored experience with program generation, verification and reuse. Applying this idea to sports reconstruction requires retaining the geometric and interaction conditions under which a repair is valid. We formulate conditional repair programs whose scene updates are checked locally and whose admission to the library requires cross-video replay and regression checks.

3 SportCraft

3.1 Overview

We organize SportCraft into three modules and two feedback loops (Figure 2). Agent-directed model fusion initializes a shared scene from complementary predictions. Knowledge-guided scene refinement forms the inner loop, diagnosing discrepancies and retaining constrained updates that pass the required checks. Agent-driven tool evolution forms the outer loop, converting development repair

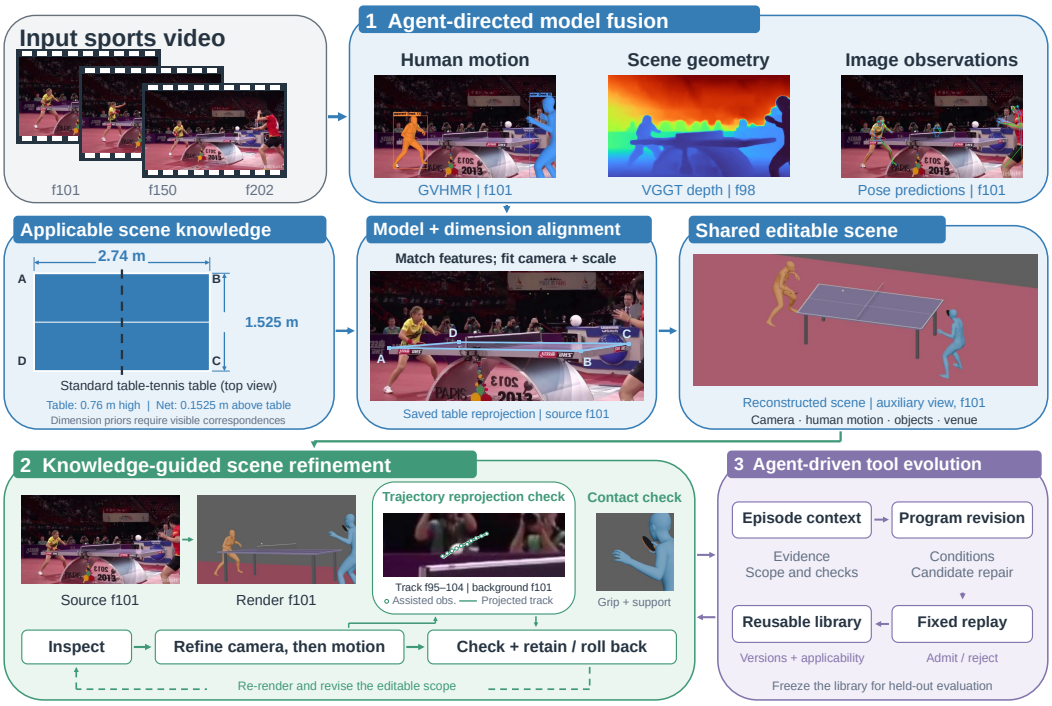

Figure 2: Overview of the proposed SportCraft framework. (1) Agent-directed model fusion registers human motion, scene geometry and image observations using venue references. (2) Knowledge-guided scene refinement uses rendered feedback and sports constraints to update the scene and check candidate changes. (3) Agent-driven tool evolution tests repair programs through replay and regression checks before reuse. Image panels illustrate these components with a table-tennis scene.

episodes into conditional procedures and checking their reuse before library admission. Pretrained model weights remain fixed.

Given video $V = \{I_t\}_{t=1}^{T}$ and sports knowledge $\mathcal{K}$, the shared scene is

$$S = (C, H, O, G, R, \Omega), \tag{1}$$

where C contains camera intrinsics and world-to-camera transforms; H human shape, articulation and root motion; O object geometry and rigid trajectories; G static venue geometry; and R time-dependent support, attachment and impact relations. The evidence record Ω binds these variables to source frames, coordinate transforms and observations, distinguishing predictions, reviewed measurements, inferences and unresolved quantities.

With memory and tools $\mathcal{H}_n = (\mathcal{M}_n, \mathcal{T}_n)$, initialization and within-video refinement produce

$$S_{n+1}^{0} = F(V_{n+1}), \qquad (S_{n+1}^{K_{\text{stop}}}, \mathcal{E}_{n+1}) = \mathcal{A}_\theta(V_{n+1}, S_{n+1}^{0}, \mathcal{K}, \mathcal{H}_n), \tag{2}$$

where n indexes development videos, K_{stop} is the inner-loop stopping iteration, and $\mathcal{E}_{n+1}$ records all attempted updates for video V_{n+1}. The outer loop uses these episodes to update memory and propose tool revisions for subsequent reconstruction tasks. Codex implements the supervisory agent $\mathcal{A}_\theta$ within the feedback loop: it inspects execution feedback, diagnoses discrepancies, proposes code changes and reviews candidate updates, coordinating the model adapters, numerical executors and scene/library transactions.

3.2 Agent-Directed Model Fusion

Independent predictors recover complementary scene components but disagree in coordinates, scale, identities and temporal support. Direct composition can therefore misplace people and equipment relative to the venue. Agent-directed model fusion (Figure 2, stage 1) registers these predictions

using supported image and venue references, while preserving provenance and leaving unsupported scale or identity assignments unresolved.

Complementary predictions. GVHMR (Shen et al., 2024) initializes human shape, articulation and root motion. VGGT (Wang et al., 2025), including the available VGGT-Ω adapter (Wang et al., 2026a), supplies camera and background-depth estimates after dynamic entities are masked. LocateAnything (Wang et al., 2026b) supplies query-conditioned boxes and derived image centers for association; it does not determine rigid pose or depth. Keypoints, masks and reviewed venue features supplement these outputs. Each prediction retains its producer, coordinate convention, preprocessing transform, timestamps and confidence type.

Shared coordinates and references. Registration inverts crop, resize and frame-sampling transforms, associates human and object identities, and registers their predictions to a common frame. A scene similarity transforms cameras, geometry, human roots and object trajectories jointly, preserving their relative errors. Cuts create separate segments unless cross-shot registration is supported; missing detections and ambiguous identities remain explicit. Identified court lines, corners, posts, goal frames or known-size objects constrain heading, camera pose and scale. Dimensions enter the fit only when the observed structure supports them; otherwise, scale remains unresolved. Uncalibrated model scores are not pooled: fitting weights depend on observation type, visibility and temporal support.

3.3 Knowledge-Guided Scene Refinement

Similar image discrepancies can arise from different errors, and improving one scene component can worsen another. Knowledge-guided scene refinement (Figure 2, stage 2) uses visual diagnosis to select editable variables and phase-specific sports constraints, orders camera, object and human updates, and retains candidates only after geometric and regression checks.

3.3.1 Diagnosis and Constraint Selection

The agent compares source frames with source-camera renders, projected geometry and residuals before choosing what to change. Coherent drift across venue references and people suggests a camera error; an isolated ball jump calls for identity, occlusion and event review; a wrist–racket discrepancy calls for local pose and contact review after camera and root checks. Auxiliary views expose relative geometry but add no independent observations.

Each diagnosis identifies the affected entities and frames, permitted variables, supporting evidence and checks that an update would invalidate. Constraints are activated only when their conditions are supported: stance can enable foot contact, flight disables it, and an observed grasp can couple hand and equipment motion. Contradicted constraints are disabled and unknown conditions remain unresolved. For selected variables z, tools instantiate

$$L(z) = \sum_{i \in \mathcal{O}_{\text{fit}}} w_i \rho(r_i^{\text{image}}(z)) + \sum_{k \in \mathcal{K}_{\text{active}}} \lambda_k \rho(r_k^{\text{knowledge}}(z)) + \lambda_{\text{anc}} \|z - z^0\|^2 + \lambda_{\text{temp}} L_{\text{temporal}}(z). \tag{3}$$

Here $\mathcal{O}_{\text{fit}}$ is the declared fitting set, $\mathcal{K}_{\text{active}} \subseteq \mathcal{K}$ the applicable sports constraints, z^0 the initial estimate and ρ a robust penalty. The weights λ_{anc} and λ_{temp} control anchoring and temporal regularization. Tools specify residual units, normalization, weights and frame support; shared observations enter once. Anchoring limits unsupported changes, while temporal terms apply within continuous interaction phases. The agent chooses the variables and applicable constraints; the numerical executor solves the resulting subproblem.

3.3.2 Geometric Updates and Acceptance

Camera, human and equipment estimates are coupled: an improved image fit can conceal a misplaced viewpoint or worsen contact elsewhere. We therefore update camera calibration before object motion and human pose, then check candidates against frozen repair criteria and regression budgets before retention.

Camera pose and scale. We refine the camera before motion so that body or object displacements cannot compensate for a misplaced viewpoint. For a world point X, projection is

$\Pi_t(X) = \pi(K_t(R_t^c X + q_t^c))$, where t indexes video frames, K_t is the intrinsic matrix, (R_t^c, q_t^c) the world-to-camera transform and π homogeneous division. Observed venue points u_{ti} correspond to 3D points X_i. The camera fit combines robust point residuals $\Pi_t(X_i) - u_{ti}$, distances to projected venue lines, the registered VGGT camera prior and within-shot temporal regularization. Planar references initialize a homography with constrained intrinsics; non-coplanar references support PnP (Theiner & Ewerth, 2023; Gutiérrez-Pérez & Agudo, 2024). Intrinsics are shared unless zoom is observed, and focal length, height and scale are not jointly released from one ambiguous plane. Positive depth, venue orientation and withheld correspondences check the result before motion is updated.

Object motion and interaction phases. With the camera fixed, an observed center defines a viewing ray; supported object size, motion and interaction geometry constrain depth. Release, flight, impact, bounce and capture divide the trajectory into intervals with different constraints. For short flight with negligible drag, $p(\tau) = p_0 + v_0(\tau - \tau_0) + \frac{1}{2}g(\tau - \tau_0)^2$, where τ is physical time, $p_0 = p(\tau_0)$ and v_0 are the initial position and velocity, and g is gravity. Physical time can differ from playback time in slow-motion footage. Metric gravity requires supported scale and capture timing. Other objects require appropriate dynamics, particularly shuttlecock drag. Fitting combines center reprojection, projected size when reliable, and event constraints. Position is continuous where appropriate, impacts may change velocity, and occluded segments remain inferred until checked at reappearance.

Human pose and equipment contact. Human refinement first adjusts root heading and translation, then selected joint rotations, using projected joints and supported contacts while anchoring the remaining pose. Foot constraints apply only during supported stance. During a grasp, the hand attachment point h_t^{hand} and an equipment-local grip point g_o contribute the residual $h_t^{\text{hand}} - T_t^o(g_o)$, where T_t^o maps equipment coordinates to the shared frame. Image alignment and contact are checked together so that improved attachment cannot conceal a worse pose. Contact changes and impacts delimit temporal regularization rather than being smoothed away.

Retention and stopping. Before optimization, the episode freezes observation roles, editable variables, repair criteria and regression budgets. A candidate is retained only if source/type invariants hold, checking evidence is sufficient, the intended repair criterion passes and every applicable local check stays within its budget. Failed candidates are recorded and rolled back; missing evidence yields an unresolved verdict. Camera changes reopen projection and depth checks, while equipment changes reopen contact and visibility checks. The loop returns the last retained scene when its criteria pass, its attempt budget expires or required evidence is unavailable.

3.4 Agent-Driven Tool Evolution

A repair that helps one video may fail when geometry, interaction phase or observation support changes. To reuse repairs under explicit conditions, agent-driven tool evolution (Figure 2, stage 3) records successes and failures as structured episodes, turns useful repairs into parameterized tools, and requires development replay and regression checks before admitting them to the reusable library.

Experience and conditional functions. Each episode stores its diagnosis, evidence, active constraints, attempted action, editable variables and before/after checks, including failures. Retrieval filters past episodes by entity type, coordinate convention, interaction phase and required support before ranking similarity. Applicability must be established again on the new video.

A useful repair can become a parameterized function with typed inputs, units, applicability conditions, dependent checks, tests and a parent version. Coordinates and identities remain inputs. For example, a grounded-stroke procedure can accept a support state and activate foot constraints only during stance. A failure can motivate changes to the solver or diagnostic tests, while scene-acceptance criteria remain fixed.

Replay and admission. Proposed tools undergo schema/unit checks, invalid-input tests, replay on new development videos and regression checks on a fixed set of earlier cases. Promotion requires improvement on targeted cases and preservation of accepted behavior within frozen tolerances:

$$\mathcal{M}_{n+1} = \mathcal{M}_n \cup \text{Distill}(\mathcal{E}_{n+1}), \qquad \mathcal{T}_{n+1} = \text{Promote}(\mathcal{T}_n, \text{ProposeTools}(\mathcal{E}_{n+1}); \mathcal{R}_n), \quad (4)$$

where $\mathcal{R}_n$ is the development replay set. Broader applicability requires evidence beyond the originating video. Parent versions support rollback, failed revisions remain recorded and retrieval excludes revoked tools. Memory and tools are frozen after development for held-out evaluation.

4 EXPERIMENTS

We select OVOW (Chen et al., 2026) and RHINO (Xue et al., 2026) as the most closely related comparison methods because both reconstruct dynamic entities together with scene geometry in a shared coordinate frame. We compare their supplied outputs with saved SportCraft scenes (Figure 3) and present a single-case qualitative ablation (Figure 4). These comparisons measure input-view reconstruction consistency and examine the effects of venue dimensions, iterative refinement and archived repair reuse.

Datasets. We evaluate 13 in-the-wild videos spanning 12 sports, with saved SportCraft scenes and supplied baselines. The sport categories and video counts are listed in Appendix A. Ground-truth 4D scenes are unavailable, precluding direct measurement of 3D geometry and motion errors.

Implementation details. Codex is integrated into the supervision loop for scene construction and refinement in SportCraft. It inspects source frames, rendered views and geometric feedback, diagnoses discrepancies, proposes code changes and reviews candidate updates against the scene checks. Model adapters supply pretrained predictions, and numerical executors solve the selected geometric subproblems. Execution results and updated renders are returned to Codex for the next review and refinement step.

Evaluation metrics. Without ground-truth 4D scenes, we assess agreement with source-derived 2D observations using three input-view consistency measures: **Human Reprojection Error (HRE)**, **Human IoU (H-IoU)** and **Object IoU (O-IoU)**, where IoU denotes intersection over union. Source-only Sapiens-2-Pose (Khirodkar et al., 2026) observations provide 12 shared body landmarks (shoulders, elbows, wrists, hips, knees and ankles), retaining confidence above 0.5. For a valid landmark j at time t, the normalized and pixel reprojection residuals are

$$e_{tj}^{\mathrm{norm}} = \frac{\|\hat{u}_{tj} - u_{tj}\|_2}{\sqrt{w_t h_t}}, \qquad e_{tj}^{\mathrm{px}} = \|\hat{u}_{tj} - u_{tj}\|_2. \tag{5}$$

Here u_{tj} is the source observation, $\hat{u}_{tj} = \Pi_t(X_{tj})$ projects the reconstructed body landmark X_{tj} with the method's final camera, and w_t, h_t are the source-observed person's bounding-box dimensions. HRE (norm) and HRE (px) average the corresponding residuals over valid joints.

Human/Object IoU compare each rendered instance mask $\hat{M}$ with its reviewed source-only SAM 3.1 (Carion et al., 2025; Meta, 2026) mask M:

$$\mathrm{IoU}(M, \hat{M}) = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|}. \tag{6}$$

Renderer ID passes supply the prediction masks. Identities and source-visible frames are fixed across methods; invisible reference instances are excluded. We average valid frames per instance, instances within each class, and sequences equally. The three measures use each method's final camera and geometry without camera refitting, ICP, Procrustes alignment or per-frame translation fitting. They assess agreement with estimated 2D observations, not independent 3D accuracy.

4.1 COMPARISON WITH EXISTING RECONSTRUCTION METHODS

The two baselines recover complementary aspects of dynamic scenes. **OVOW** (Chen et al., 2026) reconstructs instance-level mesh scenes with scale, motion, ground-contact and support constraints. **RHINO** (Xue et al., 2026) jointly models a person, a manipulated rigid object and the static scene using compositional neural fields and contact constraints. **SportCraft** represents venue geometry, articulated people and equipment as editable scene variables; its evaluated outputs are saved reconstructions. Native cameras, geometry and motion are preserved. The supplied video exports provide baseline visualization panels. Historical runtime is unavailable, so this comparison does not establish computational efficiency.

Qualitative comparisons. We compare basketball, American football and soccer reconstructions at two matched source-video times per case (Figure 3). The native views expose displaced equipment, missing actors, coarse body geometry and incomplete scene context. Venue markings in SportCraft make global placement inspectable, while comparison with the source reveals residual pose and interaction errors. The RHINO football archive has no separate object mesh. These examples show

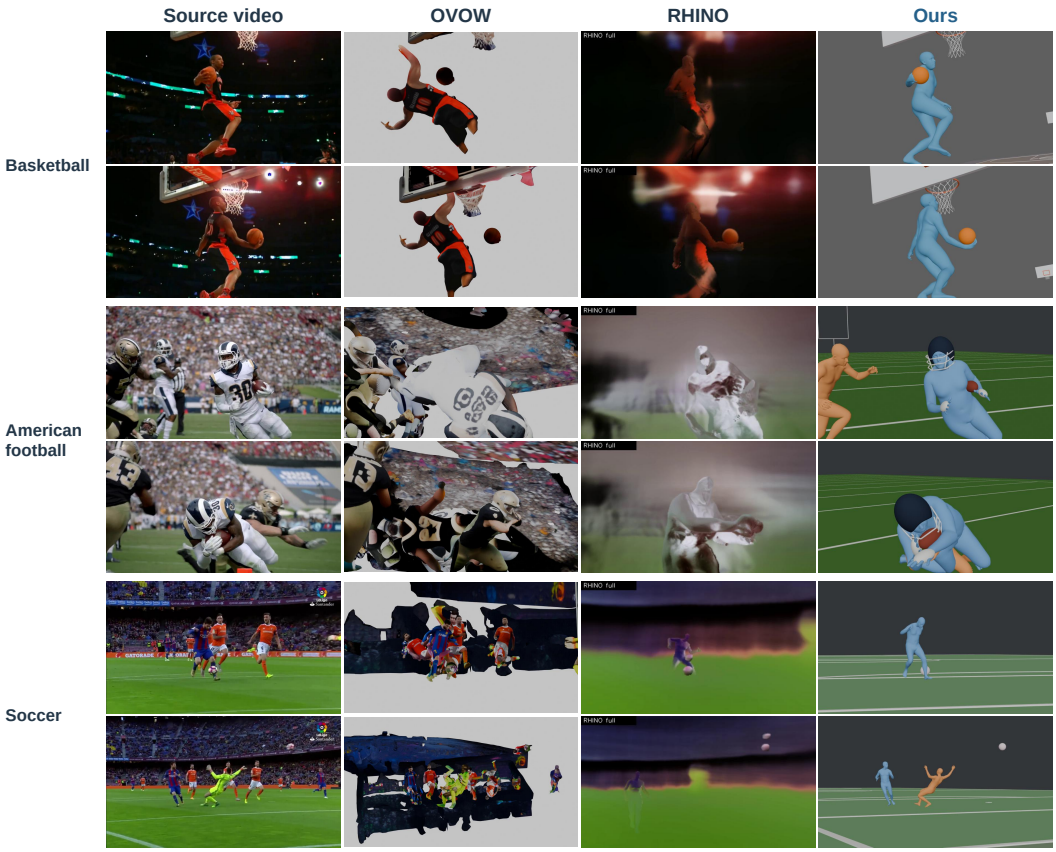

Figure 3: Qualitative comparison with OVOW and RHINO at two matched video times per case. Each sport occupies two rows in temporal order, from earlier to later. Our scenes depict athletes and equipment within explicit venue geometry. Methods use their own cameras and rendering styles. Basketball and American football show updated SportCraft scenes.

the scope and remaining errors of the delivered representations; different native cameras and rendering styles preclude interpreting appearance alone as reconstruction accuracy.

Quantitative comparisons. Our reconstructions obtain the lowest human reprojection errors and the highest Human/Object IoUs among the three methods (Table 1). OVOW outperforms RHINO on all reported measures, making it the stronger baseline in this comparison. Relative to OVOW, SportCraft reduces normalized HRE from 0.3324 to 0.0798 and pixel HRE from 180.5 to 28.2 px, corresponding to reductions of 76.0% and 84.4%, respectively. The corresponding reductions against RHINO are 86.4% and 86.0%.

Human IoU increases from 0.1261 to 0.2089 and Object IoU from 0.0649 to 0.1948 over OVOW, giving absolute gains of 0.0828 and 0.1299. Closer landmark agreement is therefore accompanied by greater silhouette overlap for both athletes and objects. The absolute IoUs remain modest, indicating substantial residual silhouette mismatch despite these gains. With each method's final camera and geometry preserved, the improvements reflect their combined agreement with source observations. These results support better input-view consistency, while depth and contact accuracy remain unverified without ground-truth 4D scenes.

4.2 Ablation Study and Discussion

We qualitatively compare three baselines to examine venue-dimension initialization, iterative refinement and reuse of an archived repair. The comparison uses a fixed refinement schedule to expose the effects of these components on scene geometry and human placement. We conduct the qualitative comparison on an ice-hockey case not used in the other manuscript figures (Figure 4).

Table 1: Quantitative comparison of input-view consistency. Human reprojection error is normalized by $\sqrt{w_t h_t}$ (norm) or reported in pixels (px); IoU values lie in $[0, 1]$. Arrows indicate the preferred direction. Bold marks the best result for each measure.

Method	HRE (norm)↓	HRE (px)↓	Human IoU↑	Object IoU↑
OVOW	0.3324	180.5	0.1261	0.0649
RHINO	0.5874	201.7	0.0555	0.0152
SportCraft	**0.0798**	**28.2**	**0.2089**	**0.1948**

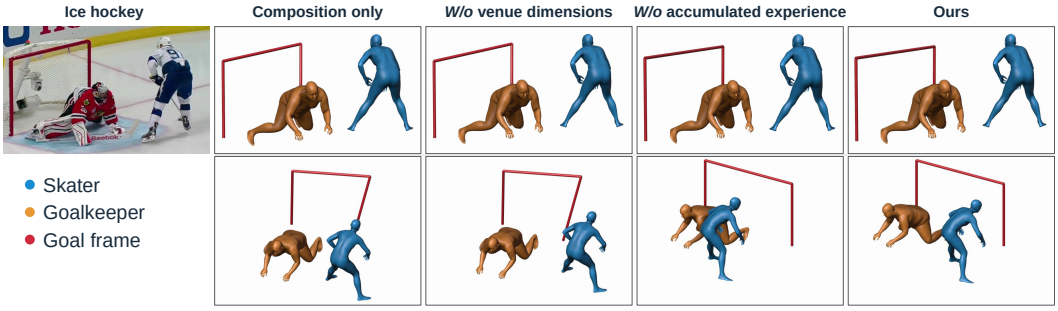

Figure 4: **Ablation study on ice hockey.** Four modeling variants at source frame 119. Top: source-camera views with a shared image crop. Bottom: auxiliary views with a shared orthographic camera and scale after rigid alignment to the goal frame, without scale normalization. Colors identify the same actors. Removing components changes the goal geometry or retained human placement; body shape and articulation remain fixed.

Composition only. We directly combine the raw model predictions and annotations after coordinate and time registration, without standard dimensions, iterative refinement or accumulated experience. Although the people and goal are assembled in one scene, the auxiliary view exposes a skewed goal frame and a different person–goal arrangement from Ours.

W/o venue dimensions. We remove the standard goal dimensions and derive its geometry from raw depth, while retaining image-based refinement and the frozen repair. The goal opening remains skewed, and the skater's position relative to the goal differs visibly from Ours. This comparison shows how dimension priors affect the inferred scene layout.

W/o accumulated experience. We disable the accumulated numerical repair and use the base solver, retaining standard dimensions and the same refinement schedule. The goal geometry is unchanged, but the skater's silhouette overlaps more with the goalkeeper's in the auxiliary view. Enabling the repair changes the retained root placement and separates their silhouettes in this view.

Discussion. The auxiliary views make the effects of goal initialization and retained human placement more apparent than the source-camera views. This comparison uses a fixed schedule and refines only root translations. The standard-dimension camera hypothesis has not passed the internal calibration criterion, so the outputs remain modeling candidates. Execution details are provided in Appendix A.

5 Conclusion

We present SportCraft, a framework for editable 4D sports reconstruction that combines agent-directed model fusion, knowledge-guided scene refinement and agent-driven tool evolution. With Codex integrated into the feedback loop, the system aligns complementary pretrained predictions, diagnoses reconstruction errors, proposes scene and tool updates, and reviews them against visual evidence and geometric checks. On in-the-wild sports videos, the evaluated reconstructions achieve closer source-view pose and mask agreement than OVOW and RHINO, while the qualitative ablation illustrates the effects of venue dimensions and archived repair reuse.

AI USE STATEMENT

AI USE STATEMENT

We used generative AI tools for implementing the proposed method and generating synthetic data used in this study. The authors reviewed and tested the code developed using these tools and performed quality checks on the generated data before use. We take full responsibility for the implementation, data, experimental results, and conclusions reported in this paper.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025.

Junhao Chen, Boran Zhang, Mingjin Chen, Henghaofan Zhang, Saining Zhang, Congcong Zhu, Hao Zhao, Ruqi Huang, Zhihao Li, and Yufei Wang. One video, one world: Turning monocular video into physical 4D scenes. In *European Conference on Computer Vision*, 2026.

Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. DreamScene4D: Dynamic multi-object scene generation from monocular videos. In *Advances in Neural Information Processing Systems*, 2024.

Thomas Gossard, Andreas Ziegler, and Andreas Zell. TT3D: Table tennis 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5867–5877, 2025.

Yunqi Gu, Ian Huang, Jihyeon Je, Guandao Yang, and Leonidas Guibas. BlenderGym: Benchmarking foundational model systems for graphics editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18574–18583, June 2025. doi: 10.1109/CVPR52734.2025.01731.

Marc Gutiérrez-Pérez and Antonio Agudo. PnLCalib: Sports field registration via points and lines optimization, 2024.

Guangzhao He, Rundong Luo, Wei-Chiu Ma, and Hadar Averbuch-Elor. Thinking in blender: Staged executable inverse graphics with vision-language models, 2026.

Ian Huang, Guandao Yang, and Leonidas Guibas. BlenderAlchemy: Editing 3D graphics with vision-language models. In *European Conference on Computer Vision*, pp. 297–314. Springer, 2024. doi: 10.1007/978-3-031-73024-5_18.

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. *arXiv preprint arXiv:2604.21681*, 2026.

Jiaxin Li, Yuxiang Wu, Zhenkai Zhang, Xinrui Shi, Haoyuan Wang, Yichen Zhao, Su Linxiang, Chenyang Yu, Mingyu Zhang, Yifan Ding, Boran Wen, Li Zhang, Ruiyang Liu, and Yong-Lu Li. HAT-4D: Lifting monocular video for 4D multi-object interactions via human-agent collaboration. In *European Conference on Computer Vision*, 2026.

Paul Liu and Jui-Hsien Wang. MonoTrack: Shuttle trajectory reconstruction from monocular badminton video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3512–3521, 2022.

Runyu Lu, Yubo Wu, Ethan Kou, Letian Fu, Wenli Xiao, Ajay Mandlekar, Yinzhen Xu, Guanya Shi, Ken Goldberg, Ang Chen, Mosharaf Chowdhury, Yuke Zhu, Linxi "Jim" Fan, and Guanzhi Wang. ASPIRE: Agentic /skills discovery for robotics. *arXiv preprint arXiv:2607.00272*, 2026. doi: 10.48550/arXiv.2607.00272.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic AI for spatial reasoning with a dynamic API. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19446–19455, June 2025.

Meta. SAM 3.1 release notes. Official SAM 3 repository release notes, 2026. Version released March 27, 2026.

Nima Rahmanian, Daniel Kienzle, Thomas Gossard, Dvij Kalaria, Rainer Lienhart, and Shankar Sastry. TT4D: A pipeline and dataset for table tennis 4D reconstruction from monocular videos, 2026.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Papers*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377.

Jonas Theiner and Ralph Ewerth. TVCalib: Camera calibration for sports field registration in soccer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1166–1175, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω, 2026a.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding, 2026b.

Shengguang Wu, Xiaohan Wang, Yuhui Zhang, Hao Zhu, and Serena Yeung-Levy. Transductive visual programming: Evolving tool libraries from experience for spatial reasoning. *arXiv preprint arXiv:2512.20934*, 2025. doi: 10.48550/arXiv.2512.20934.

Lixin Xue, Chengwei Zheng, Georgios Paschalidis, Chen Guo, Manuel Kaufmann, Juan Zarate, and Dimitrios Tzionas. RHINO: Reconstructing human interactions with novel objects from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13835–13845, June 2026.

Yixuan Yang, Zhen Luo, Wanshui Gan, Jinkun Hao, Junru Lu, Jinghao Yan, Zhaoyang Lyu, and Xudong Xu. Code-as-Room: Generating 3D rooms from top-down view images via agentic code synthesis, 2026.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21222–21232, 2023.

Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Chenyang Wang, Xiuyu Li, Michael J. Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-inverse-graphics agent via interleaved multimodal reasoning, 2026.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R. Fung, Hao Peng, and Heng Ji. CRAFT: Customizing LLMs by creating and retrieving from specialized toolsets. In *International Conference on Learning Representations*, 2024.

A Dataset and Ablation Details

Dataset composition. The evaluation set contains 13 videos across 12 sports: American football, badminton, basketball, figure skating, ice hockey, karate, kayaking, rugby, running, soccer, table tennis and volleyball. Soccer contributes two videos, and each of the remaining 11 sports contributes one.

Qualitative ablation protocol. The four conditions in Figure 4 start from the same raw VGGT/GVHMR predictions and ViTPose observations. Composition only performs the necessary timestamp, identity and coordinate registration. The dimension-enabled conditions initialize the goal with a 1.83×1.22 m hypothesis; the other conditions retain geometry inferred from raw depth. The three refinement conditions share four scheduled rounds, objectives, constraints and retention checks. Only human root translations are refined; predicted body shape and articulation remain fixed. Failed proposals retain the previous scene state.

Experience consists of an archived certified quadratic-program repair and its associated episode, applied to the unchanged numerical problem. The no-experience baseline uses the base numerical solver. Ours denotes the enabled configuration of this controlled implementation. The standard-dimension camera hypothesis remains unaccepted by the internal calibration check. This experiment uses an inspected development case to compare the geometric effects of archived repair reuse under a fixed numerical objective and refinement schedule.

The displayed example uses source frame 119 from the 50 modeled time samples spanning source frames 0–216. The full-length supplementary videos contain all 217 frames at 30 fps, with linear display interpolation between modeled samples and the saved visibility rules for unsupported actors.

B Additional Visual Results

B.1 Temporal Scene Reconstruction

Single-frame illustrations show spatial alignment but cannot convey the evolution of an interaction. We pair source frames with saved karate reconstructions at three matched times (Figure 5). Actor colors track the participants across the sequence, making their relative placement and changing poses visible. Residual contact and pose errors remain, particularly during close interactions; coarse attachment constraints do not resolve detailed hand grasping.

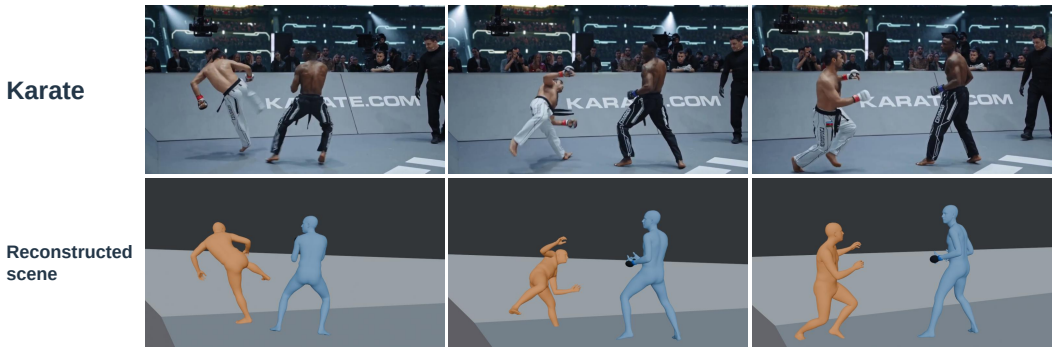

Figure 5: Temporal reconstruction of a karate interaction. Source frames (top) and our reconstruction (bottom) show the participants' changing poses and relative placement at three matched times. Colors identify the two athletes.

B.2 Side Views and Video Results

An auxiliary camera exposes spatial relationships that can be hidden in the input view. The archived basketball, American-football and soccer scenes are viewed at five source times per case (Figure 6), with one time per case shared with the baseline comparison (Figure 3). The side views reveal



Figure 6: Temporal reconstruction from source and side views. Each block shows the original video (top), the archived scene from its saved source camera (middle), and a fixed side view (bottom). Five matched times advance from left to right.

actor–ball separation and placement relative to the venue. Only the viewing camera changes; these images do not result from additional fitting.

The supplementary video collection includes source-aligned comparisons, controlled ablations and five fixed-side-view sequences covering table tennis, basketball, American football, soccer and karate. Side cameras remain stationary over each sequence to make depth motion and contact easier to inspect. The local video index provides access to these replays; unsupported observations are identified in the ablation videos. The SportCraft replays preserve missing entities and remaining reconstruction errors.