One Video, One World: Turning Monocular Video into Physical 4D Scenes

Junhao Chen[1,2⋆], Boran Zhang[3⋆], Mingjin Chen[4], Henghaofan Zhang[5], Saining Zhang[6], Congcong Zhu[3], Hao Zhao[7], Ruqi Huang[1†], Zhihao Li[2], and Yufei Wang[2†§]

[1] Shenzhen International Graduate School, Tsinghua University, China
dreamhowchen@gmail.com, ruqihuang@sz.tsinghua.edu.cn
[2] SparcAI Inc, USA
yufei.wang@sparclab.ai
[3] University of Science and Technology of China, China
[4] The Hong Kong Polytechnic University, China
[5] University of Electronic Science and Technology of China, China
[6] Nanyang Technological University, Singapore
[7] Institute for AI Industry Research (AIR), Tsinghua University, China

Project Page: https://OneVideoOneWorld.github.io

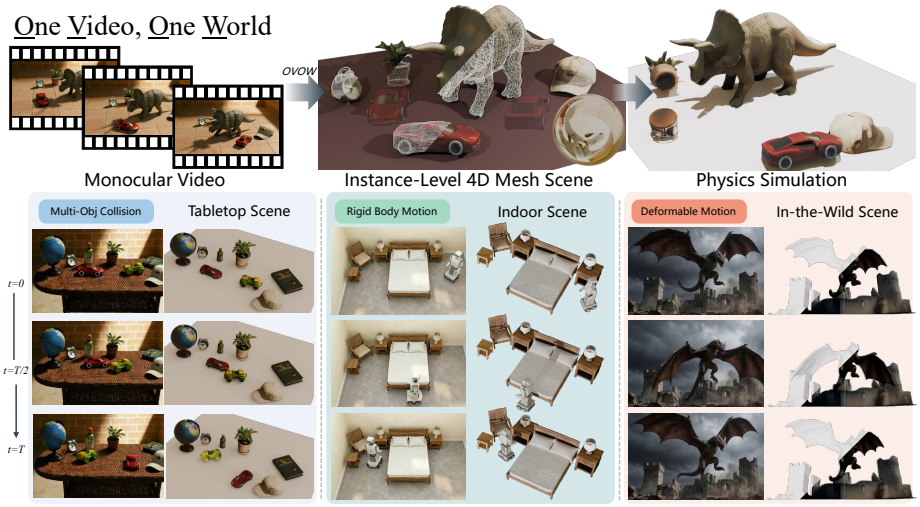

Fig. 1: OVOW reconstructs instance-level, simulation-ready 4D mesh scenes from monocular video. Given a single video, our method decomposes the scene into physically independent mesh instances and recovers rigid-body motions and non-rigid mesh deformations, yielding instance-level meshes ready for downstream physics simulation and editing. We demonstrate results of multi-object collisions, rigid-body motions, and deforming object motions across tabletop, indoor, and in-the-wild scenarios.

⋆Equal Contribution. Work performed during an internship. †Corresponding Author.
§Project Lead. Project Page at https://OneVideoOneWorld.github.io.

Abstract. We introduce **OVOW**, the first training-free system that reconstructs *instance-level, simulation-ready* 4D mesh scenes from a single monocular video. Recent 4D reconstruction achieves impressive rendering quality, but its outputs (*e.g.*, implicit fields, Gaussian primitives, or point clouds) lack the watertight topology, instance separation, and standardized physical interfaces required by physics simulators and embodied AI. OVOW closes this gap with a four-stage pipeline: a vision-language model discovers, labels, and motion-classifies all instances; category-aware reconstruction yields per-instance meshes for rigid objects and topology-consistent mesh sequences for deformable ones; an iterative render-match-optimize procedure recovers metric scale and 6-DoF pose trajectories; and physics-grounded assembly enforces ground contact and inter-object support. Crucially, we model all motion, rigid and non-rigid, through direct vertex deformation without category-specific priors or skeleton rigging, producing watertight mesh scenes ready for downstream physics simulation and editing. We further establish the first benchmark for *structured Video-to-4D* evaluation, with metrics for geometric correctness, instance separation, and physical plausibility beyond visual fidelity; the same pipeline doubles as a scalable engine for *synthesizing* paired video-to-4D simulation data for future 4D world models and embodied AI. Across two synthetic benchmarks (static and 4D), OVOW attains the best overall layout and geometry accuracy and the lowest photometric and semantic error among all baselines, and on monocular video runs one to two orders of magnitude faster than the baselines, while downstream physics simulation confirms its physical stability.

Keywords: Simulation-Ready Asset Generation · Video-to-4D · Instance-Level Scene Reconstruction

1 Introduction

Recent advances in 4D reconstruction [63,65,67,92,99] have dramatically improved the visual quality of dynamic scene rendering, yet a fundamental disconnect persists: **no existing method can produce simulation-ready 4D scene assets from video**. Physics simulators that underpin modern robotics and embodied AI, such as MuJoCo [80], Isaac Gym [57], and PyBullet [19], require watertight meshes, physically separated instances, and standardized interfaces such as URDF, which are absent from the implicit fields, Gaussian primitives, and point clouds produced by current methods [37, 61, 92]. The field excels at *looking real* but cannot yet produce *physically usable* outputs for robotic manipulation, embodied AI, gaming, and virtual reality.

Two converging trends suggest this gap can now be closed. On the *generation* side, agent-driven scene synthesis [46, 50, 87, 95, 108, 115, 121] shows growing demand for simulation-ready, instance-level assets [88], validating our target format; on the *reconstruction* side, mesh-oriented 4D [7, 35, 51] and scene-level 3D [34, 69, 104] methods show that temporally consistent mesh recovery from monocular video is feasible at the single-object level [4, 38, 116]. What remains

missing is a **unified pipeline that scales from individual objects to entire scenes**, recovering instance-level meshes with rigid and non-rigid motion and assembling them into a physically coherent, simulation-ready representation.

Achieving this requires addressing two challenges. The **first** is scene representation: how to convert a dynamic multi-object scene into simulation-compatible structured assets. Unlike skeleton-based approaches [52, 55, 112], which require predefined joint topologies and suffer from skinning artifacts on complex deformations, we adopt *direct mesh vertex deformation* as a unified motion model: each object is a mesh described by rigid-body transformations (static/rigid) or per-vertex displacement fields (deformable). This needs no predefined kinematic chains, handles articulated motion, surface deformation, and soft-body dynamics, makes *no assumptions* about object category or topology, and yields watertight, instance-level meshes ready for downstream physics simulation and editing.

The **second** challenge is data and evaluation. No dataset pairs instance-level 4D mesh annotations with source video, and no benchmark measures the structural qualities that matter for downstream physical tasks, such as geometric correctness, instance separation, and physical interaction, rather than rendering fidelity alone (PSNR/SSIM/LPIPS). We construct the *first benchmark for structured Video-to-4D evaluation*, assessing geometric and physical qualities beyond visual fidelity on carefully designed synthetic data, and further use our scalable pipeline to generate "video ↔ instance-level 4D mesh scene" pairs as data infrastructure for future 4D world-model research and embodied AI.

With these components, we present **OVOW** (Fig. 1), the first method to reconstruct instance-level, simulation-ready 4D mesh scenes from monocular video. OVOW is *fully training-free*, composing pre-trained foundation models for scene understanding, mesh reconstruction, and metric recovery into a tightly integrated pipeline, and performs robustly across diverse in-the-wild scenarios spanning indoor, outdoor, and complex multi-object scenes with rigid and non-rigid motion. On our two synthetic benchmarks for static and structured-4D scenes, OVOW achieves the best overall layout and geometry accuracy and the lowest photometric and semantic error among all baselines, and on monocular video runs one to two orders of magnitude faster than the baselines. Downstream physics simulation confirms that the reconstructed scenes remain physically stable. Our contributions are three-fold:

1. **Task & method.** We formalize *structured Video-to-4D reconstruction* and present the first training-free pipeline that turns a monocular video into instance-level, simulation-ready 4D mesh scenes, modeling *all* motion via direct vertex deformation without category priors or skeleton rigging.
2. **Benchmark.** We build the first benchmark for the task, scoring geometric correctness, instance separation, and physical plausibility beyond visual fidelity.
3. **Data engines.** We contribute two complementary data pipelines: OVOW turns real video into paired "video↔4D scene" data, and an asset-based pipeline composes the synthetic 3D/4D scenes of our benchmark; together they supply the simulation-ready paired supervision this task lacks.

2 Related Work

2.1 4D Reconstruction and Generation

3D and 4D reconstruction and generation have advanced rapidly across perception, generation, and single-object dynamics. Monocular depth estimation [47, 60, 101, 102] and feed-forward 3D prediction [81, 84, 113] recover geometry as unstructured point clouds; neural scene representations have evolved from static fields and Gaussians to dynamic 4D [3, 24, 41, 64, 119]; image/text-driven 3D generation [11, 14, 42, 54, 91, 97, 98, 114] has extended to 4D via video-diffusion priors and feed-forward models [5, 25, 43–45, 48, 53, 71]; and mesh-oriented [8, 68, 78] and motion-recovery [93, 94, 105, 111] methods recover single-object dynamic geometry. Yet these outputs remain rendering-oriented and largely single-object, lacking the watertight topology and instance separation [59,109] that physics requires. OVOW instead recovers watertight, instance-separated meshes for complete dynamic scenes, rather than single-object or rendering-only outputs.

2.2 Structured Scene Generation

A parallel line targets structured, multi-object 3D scenes, building on structured generation across text [12,27,110], image/video [10,107], and 3D [86,90]. Instance-level reconstruction [70, 79, 103, 117] recovers component-separated scenes from images, compositional generators [15–17, 30, 31, 58, 77, 83] assemble multi-object scenes, and feed-forward [32, 36, 82] and instance-aware [18] methods reconstruct multi-object motion as point clouds or Gaussians. Crucially, these build scenes from text or procedural rules, *not* real video; reconstructing structured scenes from video would unlock vast internet and robot footage [9, 13], yet existing 4D datasets [20, 94] and benchmarks [62, 120] lack source-video-paired instance-level annotations and structural metrics. In contrast, OVOW reconstructs instance-level scenes directly from real monocular video and supplies the paired data and structural benchmark this research line lacks.

2.3 Simulation-Ready Asset Generation

A growing body of work pursues assets that are ready for physical simulation. Automatic rigging [28, 29, 73, 75, 76, 100] and articulated-object generation [66, 72, 74] endow individual assets with kinematic structure such as joints and hinges for part-level articulation, while agent-driven scene generators [26, 49, 85] compose whole scenes that explicitly target simulation readiness. Together, these reflect the growing demand from physics simulators [96] and embodied world models [118] for mesh-based, physically interactive assets. However, all of these methods build assets from synthetic priors, category templates, or manual design rather than real-world visual observation. OVOW is the first to recover instance-level, simulation-ready scene meshes with both rigid and non-rigid motion directly from a monocular video, without per-category rigging or manual articulation.

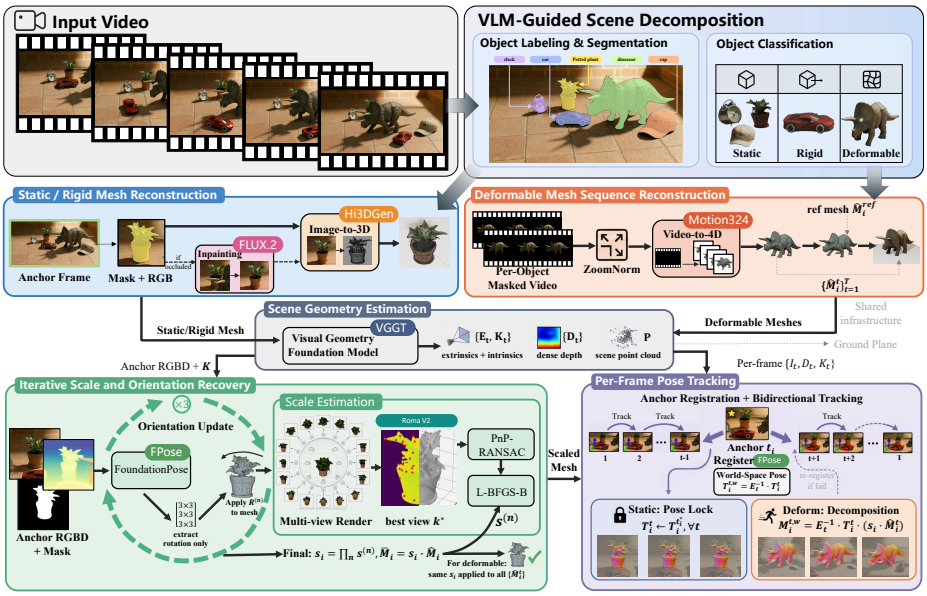

Fig. 2: Overview of OVOW (Stages 1–3). From a single monocular video, our training-free pipeline decomposes the scene into labeled, motion-classified instances, reconstructs per-instance static/rigid meshes and deformable mesh sequences, and recovers metric scale and per-frame 6-DoF motion; the recovered instances are then assembled in Fig. 3. Each stage is detailed in Sections 3.1–3.3.

3 Method

Given a monocular video $\mathcal{V} = \{I_t\}_{t=1}^{T}$ depicting a dynamic multi-object scene, our goal is to reconstruct a *simulation-ready* 4D scene

$$\mathcal{S} = \left\{ (\mathcal{M}_i,\ \{\mathbf{T}_i^t\}_{t=1}^{T}) \right\}_{i=1}^{N}, \tag{1}$$

where each of the N objects is represented by a watertight triangle mesh $\mathcal{M}_i = (\mathbf{V}_i, \mathbf{F}_i)$ with vertices $\mathbf{V}_i \in \mathbb{R}^{|\mathbf{V}_i|\times 3}$ and faces $\mathbf{F}_i$, together with a per-frame 6-DoF pose trajectory $\mathbf{T}_i^t \in \mathrm{SE}(3)$. For deformable objects, $\mathcal{M}_i$ is replaced by a topology-consistent mesh sequence $\{\mathcal{M}_i^t\}_{t=1}^{T}$ with per-vertex displacement fields. The complete scene is output as instance-level, simulation-ready meshes for downstream physics simulation and editing.

As illustrated in Figs. 2 and 3, our pipeline operates in a **fully training-free** manner and comprises four stages: **(1)** *VLM-Guided Scene Decomposition* (§3.1); **(2)** *Instance-Level Mesh Reconstruction* (§3.2); **(3)** *Spatiotemporal Pose and Deformation Recovery* (§3.3); and **(4)** *Physics-Grounded Scene Assembly* (§3.4).

3.1 VLM-Guided Scene Decomposition

We uniformly sample N_{key} keyframes (default N_{key}=3) from the input video and encode them into a multi-image prompt for a VLM [1]. The model jointly

performs *open-vocabulary object discovery*, *unique instance naming* (in the form `<noun>_<index>`), and *motion category classification* into one of three categories: **static** (unchanged pose), **rigid** (rigid-body motion), or **deformable** (non-rigid deformation). The output is a structured JSON with N records $\{(\ell_i, c_i, d_i)\}_{i=1}^{N}$, where ℓ_i is the unique label, $c_i \in \{\text{static}, \text{rigid}, \text{deform}\}$ is the motion category, and d_i is a brief visual description.

Given the label set $\{\ell_i\}$, we perform dense video segmentation using SAM3 [6] with text prompts, producing per-frame binary masks

$$\{M_i^t \in \{0,1\}^{H\times W}\}_{t=1}^{T}, \quad i = 1, \ldots, N. \tag{2}$$

Instances with maximum mask area below τ_{area} (default 200 px) are discarded.

3.2 Instance-Level Mesh Reconstruction

Static and Rigid Object Mesh Generation For each static or rigid object i ($c_i \in \{\text{static}, \text{rigid}\}$), we select the *anchor frame* $t_i^* = \arg\max_t \|M_i^t\|_1$ with the largest mask area. When the object is partially occluded, we apply an inpainting model [39] conditioned on the masked RGB, the mask, and the VLM description d_i to hallucinate the complete appearance. The completed image is then fed to a feed-forward image-to-3D model [106], producing a canonical mesh $\hat{\mathcal{M}}_i$.

Deformable Object Mesh Sequence Reconstruction For each deformable object i ($c_i = \text{deform}$), we extract a per-object masked video and apply *zoom normalization* to stabilize the object's position and scale across frames:

$$\hat{I}_i^t = \text{ZoomNorm}\big(I_t \odot M_i^t,\ \text{BBox}(M_i^t),\ \rho,\ L\big), \tag{3}$$

where $\odot$ denotes element-wise masking, ρ is the target object-to-frame ratio (default 0.65), and L is the output resolution (default 512). The normalized video is processed by a deformable mesh reconstruction model [7], which outputs a topology-consistent mesh sequence $\{\hat{\mathcal{M}}_i^t\}_{t=1}^{T}$ and a reference mesh $\hat{\mathcal{M}}_i^{\text{ref}}$.

Metric Scale Recovery We employ a visual geometry foundation model [81] to obtain per-frame camera extrinsics $\{E_t \in \text{SE}(3)\}_{t=1}^{T}$, intrinsics $\{K_t\}_{t=1}^{T}$, dense depth maps $\{D_t\}_{t=1}^{T}$, and a scene-level point cloud $\mathcal{P} \in \mathbb{R}^{P\times 3}$. For each object i, at the anchor frame t_i^*, we back-project the masked depth into 3D:

$$\mathcal{Q}_i = \Pi^{-1}\big(M_i^{t_i^*} \odot D_{t_i^*},\ K_{t_i^*}\big) \in \mathbb{R}^{Q_i\times 3}, \tag{4}$$

and compute an initial scale factor $s_i^{(0)}$ by aligning bounding-box diagonals:

$$s_i^{(0)} = \frac{\big\|\text{diag}(\mathcal{Q}_i)\big\|}{\big\|\text{diag}(\hat{\mathcal{M}}_i)\big\|}. \tag{5}$$

This estimate is refined into the final scale s_i (§3.3); the scaled mesh $\bar{\mathcal{M}}_i = s_i \cdot \hat{\mathcal{M}}_i$ is used in all subsequent stages. For deformable objects, the same s_i is applied uniformly to all mesh frames.

3.3 Spatiotemporal Pose and Deformation Recovery

We adopt a unified two-stage procedure for all object categories: *iterative scale-orientation recovery* at the anchor frame, followed by *bidirectional pose tracking* across all frames. For deformable objects, the per-frame mesh sequence replaces the single canonical mesh; otherwise the pipeline is identical.

Iterative Scale and Orientation Recovery For each object i, we jointly refine the metric scale s_i and orientation on the anchor frame t_i^* through a *render-match-optimize* loop, iterated N_{iter} times (default N_{iter}=3). At each iteration n:
(a) Orientation Update. FoundationPose [89] estimates a 6-DoF pose from the anchor-frame RGBD $(I_{t_i^*}, D_{t_i^*})$, mask $M_i^{t_i^*}$, and intrinsic $K_{t_i^*}$; we extract only the rotation $R^{(n)}$ and apply it to the mesh.
(b) Scale Estimation. The rotated mesh is rendered from J viewpoints on a sphere with z-buffer depth $\{\tilde{D}_k\}$. We establish dense correspondences between each rendered view and the real masked crop using the dense feature matcher RoMa v2 [22]. The view k^* with the most confident matches is selected, yielding 2D-3D correspondences:

$$\mathbf{q}_m^{\text{mesh}} = \mathit{\Pi}^{-1}(\mathbf{p}_m^{\text{render}},\, \tilde{D}_{k^*},\, K_{t_i^*}), \quad \mathbf{q}_m^{\text{depth}} = {R^{(n)}}^{\top} \cdot \mathit{\Pi}^{-1}(\mathbf{p}_m^{\text{real}},\, D_{t_i^*},\, K_{t_i^*}). \tag{6}$$

A pose is obtained via PnP-RANSAC [23,40], and the scalar scale $s^{(n)}$ is optimized via L-BFGS-B [2] after centering both point sets:

$$s^{(n)} = \arg\min_s \sum_m w_m \left\| \bar{\mathbf{q}}_m^{\text{depth}} - s \cdot \bar{\mathbf{q}}_m^{\text{mesh}} \right\|_2^2, \tag{7}$$

where $\bar{\mathbf{q}}$ denotes centered coordinates, w_m is the match confidence, and the top 5% residuals are trimmed. The final scale is $s_i = \prod_n s^{(n)}$ (with $s^{(0)}$=$s_i^{(0)}$ the initial estimate), yielding $\bar{\mathcal{M}}_i = s_i \cdot \hat{\mathcal{M}}_i$.

Per-Frame Pose Tracking With $\bar{\mathcal{M}}_i$, we estimate $\mathbf{T}_i^t \in \mathrm{SE}(3)$ at every frame via FoundationPose [89]:

$$\mathbf{T}_i^t = \arg\min_{\mathbf{T}} \; \mathcal{L}_{\text{render}}\big(\text{Render}(\bar{\mathcal{M}}_i, \mathbf{T}, K_t),\; I_t, D_t, M_i^t\big). \tag{8}$$

We apply *mask-gated inputs* (zeroing RGB/depth outside M_i^t) to focus registration on the target instance. At the anchor frame t_i^* the estimator performs full registration; from there we track bidirectionally:

$$\mathbf{T}_i^{t\pm1} = \texttt{Track}\big(\mathbf{T}_i^t,\; I_{t\pm1},\; D_{t\pm1},\; K_{t\pm1}\big), \tag{9}$$

falling back to full re-registration when tracking fails. The world-space pose is:

$$\mathbf{T}_i^{t,\text{w}} = E_t^{-1} \cdot \mathbf{T}_i^t. \tag{10}$$

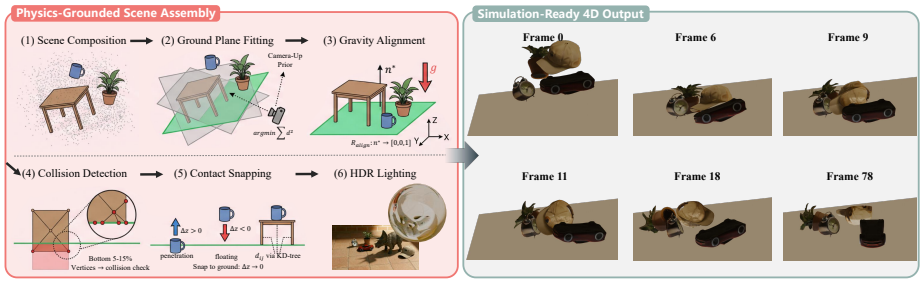

Fig. 3: Physics-Grounded Scene Assembly and Simulation-Ready Output (Stage 4). The instances from Fig. 2 are assembled into a physically coherent scene by ground-plane estimation and contact projection, then exported in URDF format for downstream physics simulation (Section 3.4).

Static Pose Locking. For *static* objects, we lock all frames to the pose at the anchor frame t_i^* (the largest-mask frame):

$$\mathbf{T}_i^t \leftarrow \mathbf{T}_i^{t_i^*}, \quad \forall\, t \neq t_i^*. \tag{11}$$

Deformable Object Decomposition. For deformable objects, the world-space mesh at frame t is:

$$\mathcal{M}_i^{t,\mathrm{w}} = \mathbf{T}_i^{t,\mathrm{w}} \cdot \bar{\mathcal{M}}_i^t = E_t^{-1} \cdot \mathbf{T}_i^t \cdot (s_i \cdot \hat{\mathcal{M}}_i^t), \tag{12}$$

decoupling global rigid trajectory from local non-rigid deformation.

3.4 Physics-Grounded Scene Assembly

The independently reconstructed objects must be assembled into a physically coherent scene. We achieve this through ground plane estimation followed by contact projection.

Ground Plane Estimation. We extract N_{plane} candidate planes $\{(\mathbf{n}_j, d_j)\}$ (default 5) from $\mathcal{P}$ via iterative RANSAC, and select the one that jointly maximizes the *above-plane ratio* $r_j = |\{\mathbf{v} : \mathbf{v}^\top \mathbf{n}_j + d_j \geq -\epsilon\}|/|\mathbf{V}_{\text{static}}|$ and minimizes the *contact proximity* $\tilde{d}_j = \text{median}(\text{bottom}_{p\%}(|\mathbf{v}^\top \mathbf{n}_j + d_j|))$ (default p=3). A *camera-up prior* resolves normal sign ambiguity:

$$\text{if } \text{median}_t\big((\mathbf{n}_j^\top R_t)_y\big) > 0, \quad \text{flip: } \mathbf{n}_j \leftarrow -\mathbf{n}_j,\ d_j \leftarrow -d_j, \tag{13}$$

where R_t is the rotation block of E_t. The selected plane is aligned to the XY-plane via R_{align} mapping $\mathbf{n}^* \to [0,0,1]^\top$.

Contact Projection. Objects are sorted by bottom distance to the ground per frame; for each object i, we extract the bottom $p_{\text{bot}}\%$ (default 10%) vertices by signed distance $d_{\mathbf{n}}(\mathbf{v}) = \mathbf{v}^\top \mathbf{n}^* + d^*$:

$$\mathcal{B}_i^t = \big\{\mathbf{v} \in \mathbf{V}_i^{t,\mathrm{w}} : d_{\mathbf{n}}(\mathbf{v}) \leq \text{quantile}_{p_{\text{bot}}}(\{d_{\mathbf{n}}(\mathbf{v}')\}_{\mathbf{v}'})\big\}. \tag{14}$$

The gravity-axis correction is:

$$\Delta z_i^t = \begin{cases} -(d_{\min} + \epsilon), & d_{\min} < -\epsilon \text{ (penetration)}, \\ -\min(d_{\min} - \epsilon,\ \delta_{\max}), & d_{\min} > \epsilon \text{ (floating)}, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where $d_{\min} = \min_{\mathbf{v} \in \mathcal{B}_i^t} d_{\mathbf{n}}(\mathbf{v})$, ϵ is a contact tolerance (0.2% of scene diagonal), and $\delta_{\max}$ caps downward settling (2%).

Inter-Object Contact. We query the nearest-surface distance from $\mathcal{B}_i^t$ to each settled object j via KD-tree:

$$d_{ij}(\mathbf{v}) = \min_{\mathbf{v}' \in \mathbf{V}_j^{t,\mathrm{w}}} \|\mathbf{v} - \mathbf{v}'\|_2, \quad \mathbf{v} \in \mathcal{B}_i^t. \tag{16}$$

If $d_{ij} \leq \tau_{\text{contact}}$ (3% of scene diagonal), the surface of j serves as a local ground plane for i, naturally handling stacking configurations. The procedure is iterated N_{asm} times (default 2) per frame to resolve cascading dependencies.

Environment Lighting Recovery. To enable photorealistic rendering of the assembled scene, we recover an HDR environment map from the input video using an intrinsic decomposition model [21], which is applied as world lighting in the exported scene.

4 Experiments

4.1 Implementation Details

OVOW composes off-the-shelf foundation models without task-specific training: Qwen3-VL [1] and SAM3 [6] for decomposition and segmentation; FLUX.2 [39] amodal inpainting with the feed-forward generator Hi3DGen [106] for static/rigid meshes; Motion324 [7] for deformable meshes; VGGT [81] for scene geometry; RoMa v2 [22] for dense correspondences; and FoundationPose [89] for 6-DoF tracking. Default settings: N_{key}=3 keyframes, τ_{area}=200 px, N_{iter}=3, N_{asm}=2, ρ=0.65, L=512, contact tolerance 0.2%, settling cap 2%, and inter-object contact threshold 3% of the scene diagonal.

4.2 Benchmark Setup

Evaluation Dataset. Separate from OVOW (which reconstructs 4D scenes from real video), we render the evaluation benchmark with a synthetic pipeline that *composes* scenes from existing 3D objects and HDRI assets in Blender, split into two complementary subsets: **OVOW-3D-Scene-Bench**, with **120** fully static scenes, and **OVOW-4D-Scene-Bench**, with **120** dynamic scenes in which at least one object undergoes rigid-body motion while the rest stay static. We sample 3D instances from Objaverse-OA [56] and compose a random **3–5** of them per scene on a ground plane with Poly Haven[1] HDRI lighting and

[1] `https://github.com/Poly-Haven/polyhavenassets`

Tab. 1: Quantitative comparison on **OVOW-3D-Scene-Bench**.

Method	Scene-IoU AABB↑	Scene-IoU OBB↑	Object IoU↑	PL ↓	N-CLIP ↓	Time (s/frame)↓	VRAM (GB)↓
CAST [104]	0.077	0.178	0.097	7.90	2.02	365	**10.5**
SAM3D [79]	0.062	0.151	0.060	9.80	2.08	**103**	33.5
VIGA [108]	**0.156**	0.158	0.024	24.40	2.61	788	31.5
MIDI [34]	0.092	0.216	0.136	8.80	1.91	234	17.7
SceneGen [58]	0.098	0.207	0.078	9.30	1.89	152	29.6
TabletopGen [87]	0.066	0.155	0.035	18.20	2.17	638	14.0
OVOW (Ours)	0.130	**0.218**	**0.190**	**5.70**	**1.87**	272	26.0

Tab. 2: Quantitative comparison on **OVOW-4D-Scene-Bench**.

Method	Scene-IoU AABB↑	Scene-IoU OBB↑	Object IoU↑	PL ↓	N-CLIP ↓	Time (s/frame)↓	VRAM (GB)↓
CAST [104]	0.091	0.176	0.080	13.10	1.74	365	**10.5**
SAM3D [79]	0.055	0.135	0.050	10.20	1.98	103	33.5
VIGA [108]	0.127	0.172	0.016	24.10	2.49	788	31.5
MIDI [34]	0.095	0.198	0.174	5.60	1.56	234	17.7
SceneGen [58]	0.096	0.183	0.051	8.20	1.54	152	29.6
TabletopGen [87]	0.072	0.200	0.034	7.50	1.91	638	14.0
OVOW (Ours)	**0.180**	**0.440**	**0.210**	**2.90**	**1.43**	**3.35**	26.0

backgrounds, animating the moving objects along procedural 6-DoF trajectories for reproducible dynamics. Every scene yields ground-truth instance meshes, 6-DoF pose trajectories, and camera parameters; full construction details are in the appendix.

Baselines. Since no existing method directly addresses instance-level 4D scene reconstruction from video, we compare against state-of-the-art single-image scene reconstruction methods: CAST [104], SAM3D [79], VIGA [108], MIDI [34], SceneGen [58], and TabletopGen [87]. All of these baselines support only single-frame image input; even on OVOW-4D-Scene-Bench they cannot consume video, so we run each independently on seven uniformly sampled frames per scene and average the resulting scores (runtime and peak VRAM are reported per generated frame), whereas OVOW consumes the full video.

Metrics. We report three IoU measures: *Scene-IoU* is the volumetric IoU of the predicted vs. ground-truth unions of object boxes, under axis-aligned (AABB↑) and oriented (OBB↑) boxes for global layout, and *Object-IoU*↑ is the per-object IoU after Hungarian matching. We also report photometric loss (PL↓, 100× normalized RGB MSE), negative-CLIP score (N-CLIP↓, 10×(1−CLIP sim)), both lower-is-better, plus per-frame time and peak VRAM.

4.3 Quantitative Comparison

Tabs. 1 and 2 report the comparison on the two benchmarks; in both, the best and second-best per column are highlighted. On OVOW-3D-Scene-Bench, OVOW obtains the best Scene-IoU-OBB (0.218), Object-IoU (0.190), PL (5.70),

and N-CLIP (1.87); VIGA attains a higher AABB-IoU but trails on every other metric, and OVOW's single-image runtime (272 s) is comparable to the feed-forward baselines. On OVOW-4D-Scene-Bench, OVOW leads every quality metric, reaching 0.440 Scene-IoU-OBB, 0.210 Object-IoU, 2.90 PL, and 1.43 N-CLIP, while amortizing computation across the video to run at 3.35 s per frame, one to two orders of magnitude faster than the baselines (103–788 s).

Where the gains come from. OVOW leads on both benchmarks, with the largest margin on the dynamic 4D scenes, where video-level temporal cues are most informative. Amodal inpainting keeps per-object geometry faithful under occlusion, and physics-grounded assembly resolves ground contact and inter-object support, so scenes show far fewer interpenetration and floating artifacts than the generative baselines and stay stable under gravity in a physics simulator.

Scalability. Performance degrades gracefully as the number of objects per scene grows; the mild quality drop on the most crowded scenes is primarily caused by increased mutual occlusion among instances, while per-frame runtime grows only modestly.

4.4 Qualitative Comparison

Fig. 4 compares OVOW against feed-forward 3D reconstruction (VGGT, Depth Anything 3) and instance-level scene-reconstruction baselines on a dynamic scene: OVOW uniquely recovers instance-separated, watertight meshes with accurate layout, while the feed-forward methods lack instance separation and the scene-reconstruction baselines misplace objects or distort geometry. Fig. 5 shows OVOW applied to single images (image-to-3D): from one in-the-wild frame it recovers instance-separated meshes with faithful shapes and accurate layout through iterative scale-orientation refinement, with per-baseline comparisons deferred to the appendix. OVOW is the only method that simultaneously achieves accurate geometry, correct spatial layout, and watertight, simulation-ready topology, benefiting from video-level temporal consistency for scale and pose recovery and from amodal inpainting for occluded instances. Fig. 7 further contrasts OVOW with two tempting alternatives: single-object video-to-mesh generation (DreamScene4D) degrades at unseen viewpoints and recovers neither metric scale nor inter-object contact, while depth-fusion (Depth Anything 3 followed by NKSR) yields a single fused surface without instance separation. Beyond per-scene reconstruction, Fig. 6 shows paired video-to-4D examples that OVOW generates across diverse scene types, illustrating the data our pipeline can produce at scale. Additional per-scene visualizations and comparisons are provided in the appendix.

4.5 Sensitivity of Key Design Choices

Hyperparameter robustness. On a held-out validation set spanning static, rigid, and deformable motion, OVOW is non-brittle to its key design choices (Tab. 3). Refinement iterations are the accuracy–runtime elbow (IoU-B saturates

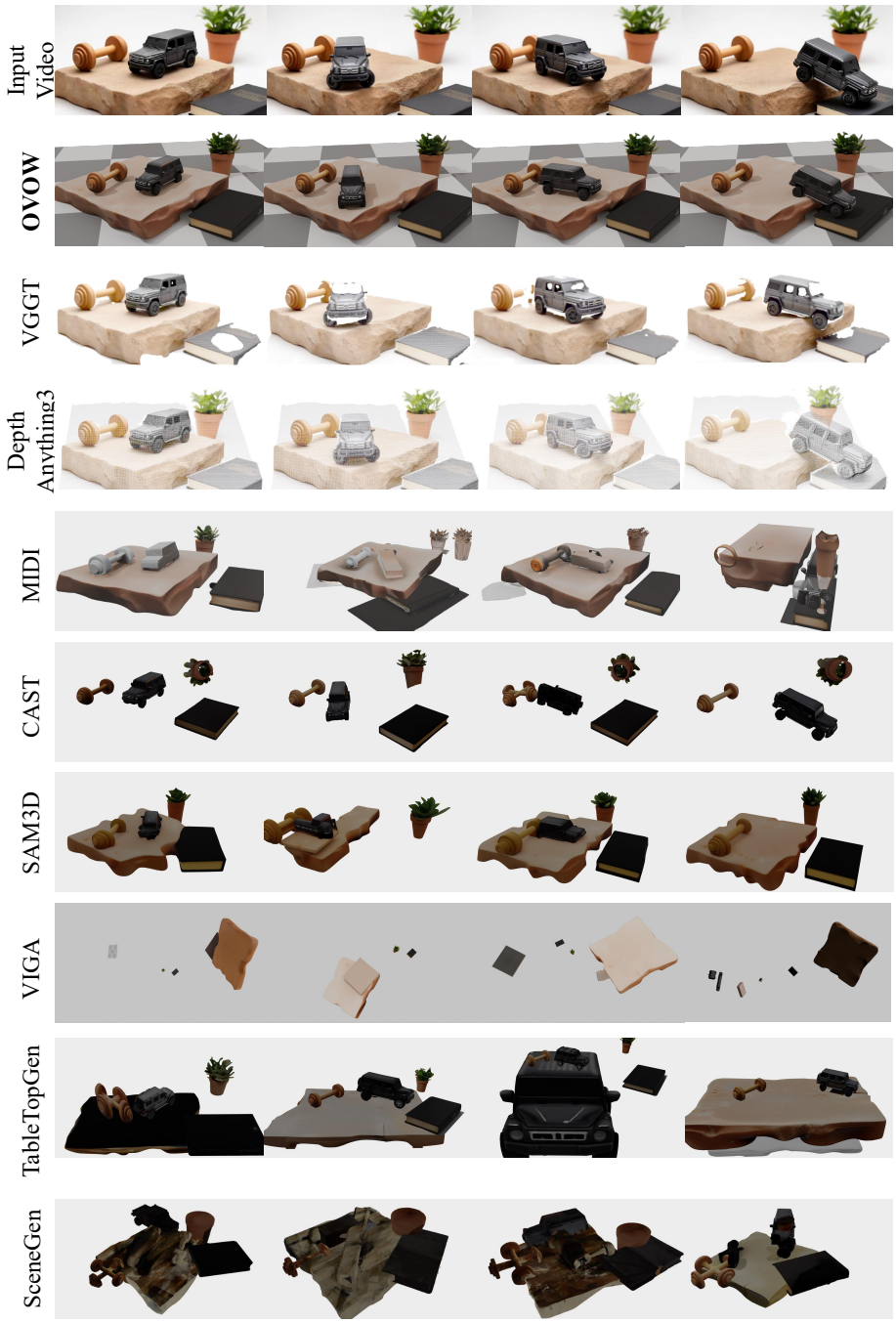

Fig. 4: Qualitative comparison on an in-the-wild dynamic scene. Four frames of a dynamic scene comparing OVOW against feed-forward 3D reconstruction (VGGT [81], Depth Anything 3 [47]) and instance-level scene-reconstruction methods.

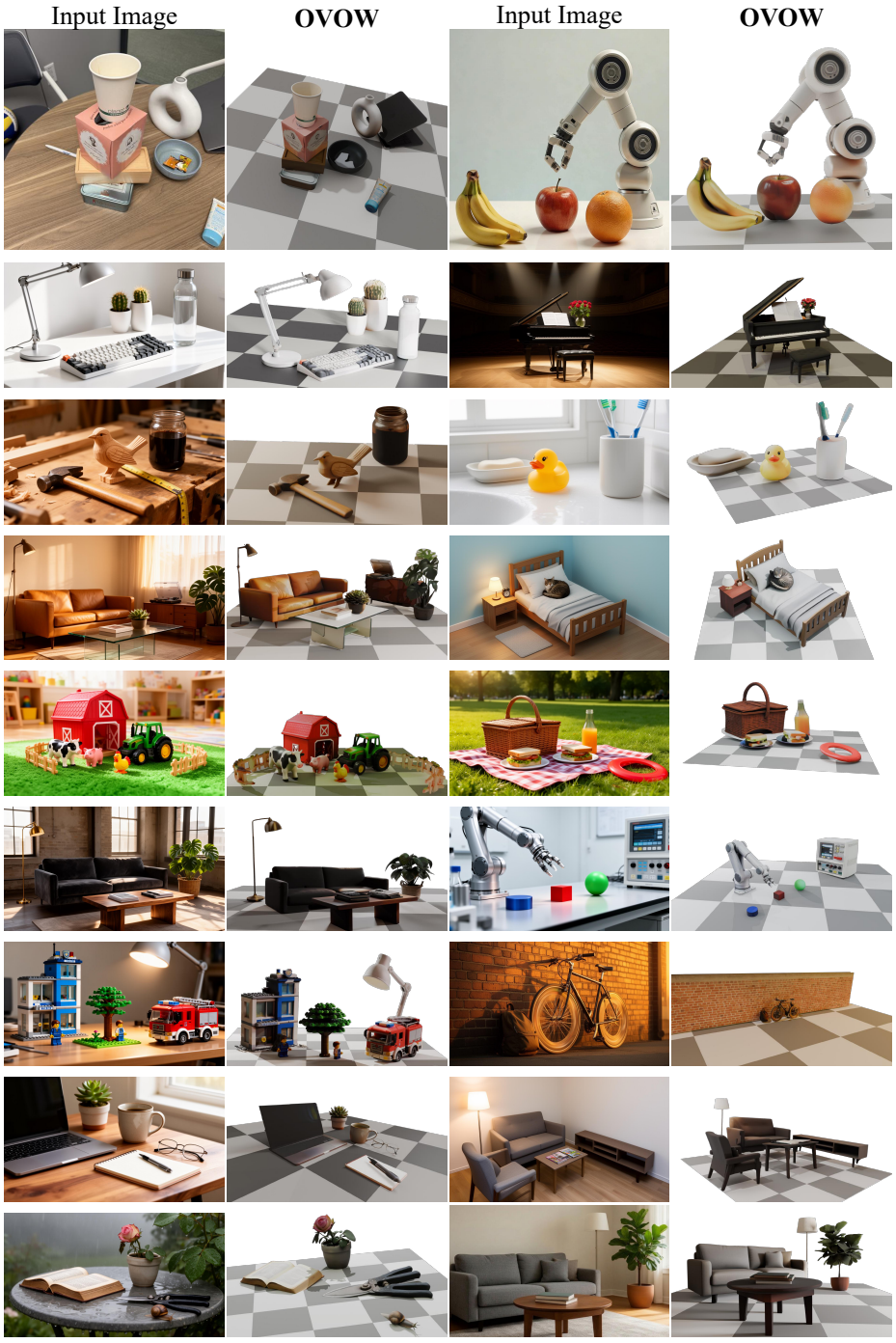

Fig. 5: Image-to-3D reconstruction by OVOW.

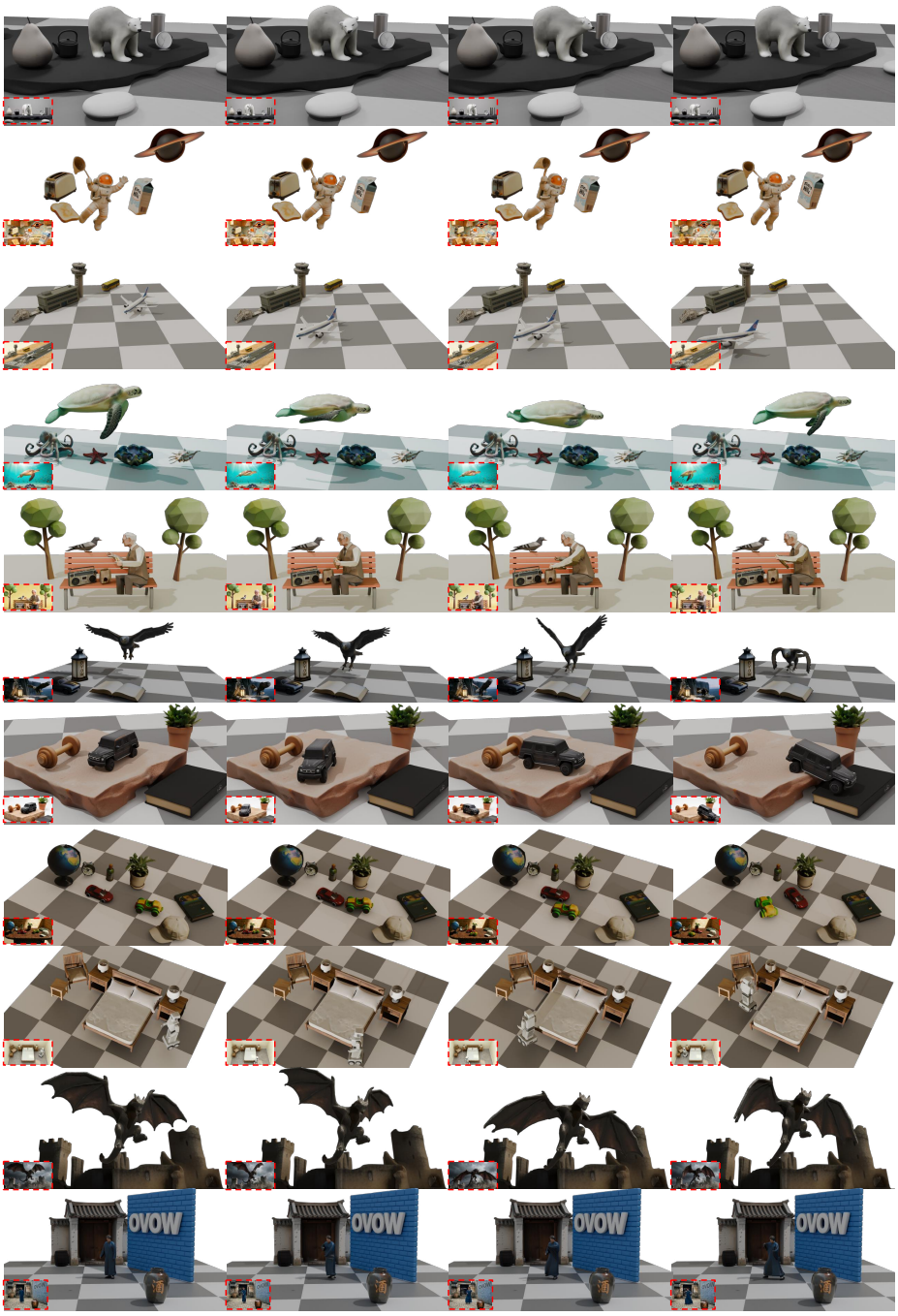

Fig. 6: Paired video-to-4D results generated by OVOW. Each example is a reconstructed instance-level 4D mesh scene; the red box at the bottom-left corner of each render shows the corresponding input frame.

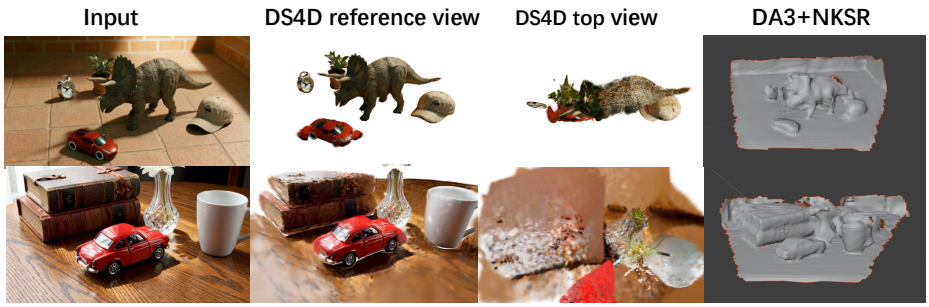

Fig. 7: Comparison with video-to-mesh and fused-surface alternatives. DreamScene4D [18] (DS4D) is plausible at the reference view but degrades at other viewpoints, while Depth Anything 3 [47] followed by NKSR [33] (DA3+NKSR) fuses a single surface without instance separation, metric scale, or contact. OVOW instead outputs instance-separated, simulation-ready meshes.

at 0.78 by N_{iter}=3); ρ=0.65 and τ_{area}=200 px are near-optimal; and the contact parameters (ϵ=0.2%, $\delta_{\max}$=2%, p_{bot}=10%) share τ_{contact}'s flat stability plateau.

Tab. 3: Hyperparameter robustness on the validation set (spanning static, rigid, and deformable motion). Defaults are underlined; "–" marks unused cells. Metrics are bounding-box IoU (IoU-B), valid-scene rate (%), and simulation stability (%).

Hyperparameter	Metric	Values (default underlined)			
$N_{\text{iter}} \in \{1, 2, \underline{3}, 5\}$	IoU-B↑	0.58	0.72	**0.78**	0.79
$\rho \in \{0.55, \underline{0.65}, 0.75\}$	IoU-B↑	0.74	**0.78**	0.76	–
$\tau_{\text{area}} \in \{100, \underline{200}, 400\}$	Valid scenes↑	84.9	**86.8**	85.7	–
$\tau_{\text{contact}} \in \{2, \underline{3}, 4\}\%$	Sim. stability↑	79.8	**82.7**	81.5	–
$N_{\text{asm}} \in \{1, \underline{2}, 5\}$	Sim. stability↑	78.9	**82.7**	82.8	–

Stage-level reliability. The pipeline stages are also reliable on this validation set: 95.4% motion-category accuracy, 93.1%/88.7% rigid/deformable reconstruction success, 92.4% pose recovery, and 86.8% final validity with 82.7% simulation stability, so failures rarely propagate.

5 Conclusion

We presented OVOW, the first training-free system that turns monocular video into instance-level, simulation-ready 4D mesh scenes, modeling all motion via direct vertex deformation without rigging or category priors. It leads the baselines in geometry, layout, and photometric/semantic accuracy, runs one to two orders of magnitude faster on video, and doubles as a scalable engine for paired video-to-4D simulation data.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM Journal on scientific computing **16**(5), 1190–1208 (1995)
3. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
4. Cao, Y., Lu, J., Huang, Z., Shen, Z., Zhao, C., Hong, F., Chen, Z., Li, X., Wang, W., Liu, Y., et al.: Reconstructing 4d spatial intelligence: A survey. arXiv preprint arXiv:2507.21045 (2025)
5. Cao, Z., Hong, F., Chen, Z., Pan, L., Liu, Z.: Physx-anything: Simulation-ready physical 3d assets from single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5839–5848 (2026)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
7. Chen, H., Chen, X., Xu, Z., Chen, A.: Motion 3-to-4: 3d motion reconstruction for 4d synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28947–28958 (2026)
8. Chen, J., Zhang, B., Tang, X., Wonka, P.: V2m4: 4d mesh animation reconstruction from a single monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11643–11653 (2025)
9. Chen, J., Chen, M., Xu, J., Li, X., Dong, J., Sun, M., Jiang, P., Li, H., Yang, Y., Zhao, H., Long, X.X., Huang, R.: Dancetogether: Generating interactive multi-person video without identity drifting. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=7VEECFBzmm
10. Chen, J., Gao, K., Cui, Y., Sun, M., Chen, M., Wang, S., Long, X., Ma, F., Tian, Q., Zhao, H., Huang, R.: Lottiegpt: Tokenizing vector animation for autoregressive generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 31639–31651 (June 2026)
11. Chen, J., Li, X., Ye, X., Li, C., Fan, Z., Zhao, H.: Idea23d: Collaborative lmm agents enable 3d model generation from interleaved multimodal inputs. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 4149–4166 (2025)
12. Chen, J., Sun, J., Li, X., Xin, H., Xue, Y., Xu, Y., Zhao, H.: LLMsPark: A benchmark for evaluating large language models in strategic gaming contexts. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 182–194. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.12, https://aclanthology.org/2025.findings-emnlp.12/
13. Chen, M., Chen, J., Fan, Z., Lee, Y., Dang, Z., Wang, L., Cui, Y., Chau, L.P., Wang, Y.: Hvg-3d: Bridging real and simulation domains for 3d-conditional hand-object interaction video synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15986–15997 (June 2026)

14. Chen, M., Chen, J., Gao, H.a., Chen, X., Fan, Z., Zhao, H.: Ultraman: ultra-fast and high-resolution texture generation for 3d human reconstruction from a single image. Machine Vision and Applications **37**(2), 24 (2026)
15. Chen, M., Wang, L., Ao, S., Zhang, Y., Xu, K., Guo, Y.: Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors (2025), https://arxiv.org/abs/2501.02519
16. Chen, M., Yang, R., Hu, Q., Xue, K., Zhou, S., Guo, Y.: Graph2scene: Versatile 3d indoor scene generation with interaction-aware scene graph. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11313–11320 (2025). https://doi.org/10.1109/IROS60139.2025.11246595
17. Chen, W., Rao, J., Wang, W., Li, X., Cheng, X., Cao, L.: Customtex: High-fidelity indoor scene texturing via multi-reference customization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4280–4290 (June 2026)
18. Chu, W.H., Ke, L., Fragkiadaki, K.: Dreamscene4d: Dynamic multi-object scene generation from monocular videos. Advances in Neural Information Processing Systems **37**, 96181–96206 (2024)
19. Coumans, E., Bai, Y.: Pybullet, a python module for physics simulation for games, robotics and machine learning. https://pybullet.org/wordpress/ (2016–2021)
20. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
21. Dille, S., Careaga, C., Aksoy, Y.: Intrinsic single-image hdr reconstruction. In: European Conference on Computer Vision. pp. 161–177. Springer (2024)
22. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching. arXiv preprint arXiv:2511.15706 (2025)
23. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
24. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12479–12488 (2023)
25. Geng, Z., Wang, N., Xu, S., Ye, C., Li, B., Chen, Z., Peng, S., Zhao, H.: One view, many worlds: Single-image to 3d object meets generative domain randomization for one-shot 6d pose estimation. In: 9th Annual Conference on Robot Learning (2025), https://openreview.net/forum?id=kto4zVmo4w
26. Gu, Z., Cui, Y., Li, Z., Wei, F., Ge, Y., Gu, J., Liu, M.Y., Davis, A., Ding, Y.: Artiscene: Language-driven artistic 3d scene generation through image intermediary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2891–2901 (June 2025)
27. Guo, H., Zhang, W., Chen, J., Gu, Y., Yang, J., Du, J., Cao, S., Hui, B., Liu, T., Ma, J., Zhou, C., Li, Z.: IW-bench: Evaluating large multimodal models for converting image-to-web. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 6449–6466. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.334, https://aclanthology.org/2025.findings-acl.334/

28. Guo, J., Liu, J., Chen, J., Mao, S., Hu, C., Jiang, P., Yu, J., Xu, J., Liu, Q., Xu, L., et al.: Auto-connect: Connectivity-preserving rigformer with direct preference optimization. arXiv preprint arXiv:2506.11430 (2025)
29. Guo, Z., Zhang, O., Xiang, J., Zhao, A., Zhou, W., Li, H.: Make-it-poseable: Feed-forward latent posing model for 3d humanoid character animation. arXiv preprint arXiv:2512.16767 (2025)
30. Han, H., Yang, R., Liao, H., Xing, J., Xu, Z., Yu, X., Zha, J., Li, X., Li, W.: Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25367–25377 (2025)
31. Hu, S., Arroyo, D.M., Debats, S., Manhardt, F., Carlone, L., Tombari, F.: Mixed diffusion for 3d indoor scene synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1262–1272 (2026)
32. Hu, Y., Yang, Y., Lin, H., Wang, Y., Dong, J., Deng, Y., Zhu, X., Jia, F., Bao, H., Zhou, X., et al.: Split4d: Decomposed 4d scene reconstruction without video segmentation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
33. Huang, J., Gojcic, Z., Atzmon, M., Litany, O., Fidler, S., Williams, F.: Neural kernel surface reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4369–4379 (2023)
34. Huang, Z., Guo, Y.C., An, X., Yang, Y., Li, Y., Zou, Z.X., Liang, D., Liu, X., Cao, Y.P., Sheng, L.: Midi: Multi-instance diffusion for single image to 3d scene generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23646–23657 (2025)
35. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Mesh4d: 4d mesh reconstruction and tracking from monocular video. arXiv preprint arXiv:2601.05251 (2026)
36. Karhade, J., Keetha, N., Zhang, Y., Gupta, T., Sharma, A., Scherer, S., Ramanan, D.: Any4d: Unified feed-forward metric 4d reconstruction. arXiv preprint arXiv:2512.10935 (2025)
37. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023)
38. Kong, L., Yang, W., Mei, J., Liu, Y., Liang, A., Zhu, D., Lu, D., Yin, W., Hu, X., Jia, M., et al.: 3d and 4d world modeling: A survey. arXiv preprint arXiv:2509.07996 (2025)
39. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2` (2025)
40. Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. International journal of computer vision **81**(2), 155–166 (2009)
41. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6498–6508 (2021)
42. Li, Z., Wang, Y., Zheng, H., Luo, Y., Wen, B.: Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), `https://openreview.net/forum?id=yslRXs9gcJ`
43. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. Advances in Neural Information Processing Systems **37**, 21377–21400 (2024)

44. Liang, H., Xu, D., Bhatt, N.P., Hu, H., Liang, H., Plataniotis, K.N.: Comp4d: Compositional 4d scene generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3567–3577 (2026)
45. Liang, H., Yin, Y., Xu, D., Liang, H., Wang, Z., Plataniotis, K.N., Zhao, Y., Wei, Y.: Diffusion4d: fast spatial-temporal consistent 4d generation via video diffusion models. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 110854–110875 (2024)
46. Lin, G., Huang, K., Liu, M., Gao, R., Chen, H., Chen, L., Lu, B., Komura, T., Liu, Y., Zhu, J.Y., et al.: Pat3d: Physics-augmented text-to-3d scene generation. arXiv preprint arXiv:2511.21978 (2025)
47. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
48. Lin, J., Wang, Z., Xu, D., Jiang, S., Gong, Y., Jiang, M.: Phys4dgen: Physics-compliant 4d generation with multi-material composition perception. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10398–10407 (2025)
49. Ling, L., Ge, Y., Sheng, Y., Bera, A.: I-scene: 3d instance models are implicit generalizable spatial learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26974–26983 (2026)
50. Ling, L., Lin, C.H., Lin, T.Y., Ding, Y., Zeng, Y., Sheng, Y., Ge, Y., Liu, M.Y., Bera, A., Li, Z.: Scenethesis: A language and vision agentic framework for 3d scene generation. arXiv preprint arXiv:2505.02836 (2025)
51. Liu, I., Su, H., Wang, X.: Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=LuGHbK8qTa
52. Liu, I., Xu, Z., Yifan, W., Tan, H., Xu, Z., Wang, X., Su, H., Shi, Z.: Riganything: Template-free autoregressive rigging for diverse 3d assets. ACM Transactions on Graphics (TOG) **44**(4), 1–12 (2025)
53. Liu, T., Huang, Z., Chen, Z., Wang, G., Hu, S., Shen, L., Sun, H., Cao, Z., Li, W., Liu, Z.: Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25571–25582 (2025)
54. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
55. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (Oct 2015)
56. Lu, Y., Tian, Y., Jiang, Z., Zhao, Y., Yang, Y., Ouyang, H., Hu, H., Yu, H., Shen, Y., Liao, Y.: Orientation matters: Making 3d generative models orientation-aligned. Advances in Neural Information Processing Systems **38**, 148548–148583 (2026)
57. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
58. Meng, Y., Wu, H., Zhang, Y., Xie, W.: Scenegen: Single-image 3d scene generation in one feedforward pass. In: Thirteenth International Conference on 3D Vision (2026), https://openreview.net/forum?id=KpelxFFxQT

59. Miao, Q., Li, K., Quan, J., Min, Z., Ma, S., Xu, Y., Yang, Y., Liu, P., Luo, Y.: Advances in 4d generation: A survey. arXiv preprint arXiv:2503.14501 (2025)
60. Miao, X., Dong, J., Zhao, Q., Yang, Y., Chen, J., Long, Y.: From frames to sequences: Temporally consistent human-centric dense prediction (2026), https://arxiv.org/abs/2602.01661
61. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
62. Nazarczuk, M., Tanay, T., Moreau, A., Zhang, Z., Pérez-Pellitero, E.: Charge: A comprehensive novel view synthesis benchmark and dataset to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15323–15333 (2026)
63. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021)
64. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics (TOG) **40**(6), 1–12 (2021)
65. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10318–10327 (2021)
66. Qiu, X., Yang, J., Wang, Y., Chen, Z., Wang, Y., Wang, T.H., Xian, Z., Gan, C.: Articulate anymesh: Open-vocabulary 3d articulated objects modeling. In: 9th Annual Conference on Robot Learning (2025), https://openreview.net/forum?id=BNCh3SS1Yl
67. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. Advances in Neural Information Processing Systems **37**, 56828–56858 (2024)
68. Sabathier, R., Novotny, D., Mitra, N.J., Monnier, T.: Actionmesh: Animated 3d mesh generation with temporal 3d diffusion. arXiv preprint arXiv:2601.16148 (2026)
69. Shi, Y., Li, W., Wang, Z., Li, H., Chen, X., Tan, P., Zhang, L.: Scenemaker: Open-set 3d scene generation with decoupled de-occlusion and pose estimation model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27146–27156 (2026)
70. Siddiqui, Y., Frost, D., Aroudj, S., Avetisyan, A., Howard-Jenkins, H., DeTone, D., Moulon, P., Wu, Q., Li, Z., Straub, J., et al.: Shaper: Robust conditional 3d shape generation from casual captures. arXiv preprint arXiv:2601.11514 (2026)
71. Singer, U., Sheynin, S., Polyak, A., Ashual, O., Makarov, I., Kokkinos, F., Goyal, N., Vedaldi, A., Parikh, D., Johnson, J., et al.: Text-to-4d dynamic scene generation. In: Proceedings of the 40th International Conference on Machine Learning. pp. 31915–31929 (2023)
72. Song, C., Li, X., Yang, F., Xu, Z., Wei, J., Liu, F., Feng, J., Lin, G., Zhang, J.: Puppeteer: Rig and animate your 3d models. Advances in neural information processing systems **38**, 72152–72184 (2026)
73. Song, C., Zhang, J., Li, X., Yang, F., Chen, Y., Xu, Z., Liew, J.H., Guo, X., Liu, F., Feng, J., et al.: Magicarticulate: Make your 3d models articulation-ready. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15998–16007 (2025)

74. Song, C., Zhang, J., Li, X., Yang, F., Chen, Y., Xu, Z., Liew, J.H., Guo, X., Liu, F., Feng, J., et al.: Magicarticulate: Make your 3d models articulation-ready. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15998–16007 (2025)
75. Sun, M., Chen, J., Dong, J., Chen, Y., Jiang, X., Mao, S., Jiang, P., Wang, J., Dai, B., Huang, R.: Drive: Diffusion-based rigging empowers generation of versatile and expressive characters. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21170–21180 (2025)
76. Sun, M., Zeng, C., Pei, J., Chen, J., Song, C., Wang, S., Chang, T., Huang, B., Zeng, Z., Huang, R.: Animator-centric skeleton generation on objects with fine-grained details. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17336–17345 (June 2026)
77. Tang, X., Li, R., Fan, X.: Zeroscene: A zero-shot framework for 3d scene generation from a single image and controllable texture editing. In: Computer Graphics Forum. p. e70419. Wiley Online Library (2026)
78. Tao, S., Zhou, B., Tu, H., Wang, Y., Liu, Y.: Tesselation gs: Neural mesh gaussians for robust monocular reconstruction of dynamic objects. In: Thirteenth International Conference on 3D Vision (2025)
79. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025), https://arxiv.org/abs/2511.16624
80. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ international conference on intelligent robots and systems. pp. 5026–5033. IEEE (2012)
81. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
82. Wang, J., Che, H., Chen, Y., Yang, Z., Goli, L., Manivasagam, S., Urtasun, R.: Flux4d: Flow-based unsupervised 4d reconstruction. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=FeUGQ6AiKR
83. Wang, L., Guo, H.X., Wang, X., Sun, F., Sun, K., Liu, P., Xiao, H., Wang, Z., Fu, G., Li, E., Liu, Y., Wang, Y.: Scenetransporter: Optimal transport-guided compositional latent diffusion for single-image structured 3d scene generation. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=xjCkwPhQWq
84. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20697–20709 (2024)
85. Wang, X., Liu, L., Cao, Y., Wu, R., Qin, W., Wang, D., Sui, W., Su, Z.: Embodiedgen: Towards a generative 3d world engine for embodied intelligence. arXiv preprint arXiv:2506.10600 (2025)
86. Wang, Z., Lorraine, J., Wang, Y., Su, H., Zhu, J., Fidler, S., Zeng, X.: Llama-mesh: Unifying 3d mesh generation with language models. arXiv preprint arXiv:2411.09595 (2024)
87. Wang, Z., He, Y., Yang, L., Zou, W., Ma, H., Liu, L., Sui, W., Guo, Y., Su, H.: Tabletopgen: Instance-level interactive 3d tabletop scene generation from text or single image. arXiv preprint arXiv:2512.01204 (2025)
88. Wen, B., Xie, H., Chen, Z., Hong, F., Liu, Z.: 3d scene generation: A survey. arXiv preprint arXiv:2505.05474 (2025)

89. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17868–17879 (2024)
90. Weng, F., Chen, J., Li, X., Qin, J., Guo, H., ShaochunHao, Han, X.: GarmentGPT: Compositional garment pattern generation via discrete latent tokenization. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=XzXKnazRBF
91. Weng, J., Zhang, S., Diao, Z., Li, P., Zhang, H., Chen, J., Zhao, H.: Feedforward 3d editing learns from semantic-part transformation. arXiv preprint arXiv:2605.27351 (2026)
92. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
93. Wu, Y., Chen, Z., Liu, S., Ren, Z., Wang, S.: Casa: Category-agnostic skeletal animal reconstruction. Advances in Neural Information Processing Systems **35**, 28559–28574 (2022)
94. Wu, Z., Yu, C., Wang, F., Bai, X.: Animateanymesh: A feed-forward 4d foundation model for text-driven universal mesh animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13557–13568 (2025)
95. Xia, H., Li, X., Li, Z., Ma, Q., Xu, J., Liu, M.Y., Cui, Y., Lin, T.Y., Ma, W.C., Wang, S., et al.: Sage: Scalable agentic 3d scene generation for embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22358–22368 (2026)
96. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11097–11107 (2020)
97. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., et al.: Native and compact structured latents for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14419–14429 (2026)
98. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
99. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: SV4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=tJoS2d0Onf
100. Xu, Z., Zhou, Y., Kalogerakis, E., Landreth, C., Singh, K.: Rignet: Neural rigging for articulated characters. ACM transactions on graphics **39**(4) (2020)
101. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
102. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
103. Yang, Z., Yang, B., Dong, W., Cao, C., Cui, L., Ma, Y., Cui, Z., Bao, H.: Instascene: Towards complete 3d instance decomposition and reconstruction from cluttered scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7771–7781 (2025)

104. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. ACM Transactions on Graphics (TOG) **44**(4), 1–19 (2025)
105. Yao, Y., Deng, Z., Hou, J.: Riggs: Rigging of 3d gaussians for modeling articulated objects in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5592–5601 (2025)
106. Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., Zhao, H., Han, X.: Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25050–25061 (2025)
107. Ye, X., Chen, J., Li, X., Xin, H., Li, C., Zhou, S., Bu, J.: Mmad: Multi-modal movie audio description. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11415–11428 (2024)
108. Yin, S., Ge, J., Wang, Z.Z., Li, X., Black, M.J., Darrell, T., Kanazawa, A., Feng, H.: Vision-as-inverse-graphics agent via interleaved multimodal reasoning. arXiv preprint arXiv:2601.11109 (2026)
109. Yunus, R., Lenssen, J.E., Niemeyer, M., Liao, Y., Rupprecht, C., Theobalt, C., Pons-Moll, G., Huang, J.B., Golyanik, V., Ilg, E.: Recent trends in 3d reconstruction of general non-rigid scenes. In: Computer Graphics Forum. vol. 43, p. e15062. Wiley Online Library (2024)
110. Zhang, B., Xie, H., Du, P., Chen, J., Cao, P., Chen, Y., Liu, S., Liu, K., Zhao, J.: Zhujiu: A multi-dimensional, multi-faceted chinese benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 479–494 (2023)
111. Zhang, H., Chang, D., Li, F., Soleymani, M., Ahuja, N.: Magicpose4d: Crafting articulated models with appearance and motion control. Transactions on Machine Learning Research **2025** (2025)
112. Zhang, J.P., Pu, C.F., Guo, M.H., Cao, Y.P., Hu, S.M.: One model to rig them all: Diverse skeleton rigging with unirig. ACM Transactions on Graphics (TOG) **44**(4), 1–18 (2025)
113. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: MonST3r: A simple approach for estimating geometry in the presence of motion. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=lJpqxFgWCM
114. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. ACM Transactions on Graphics (TOG) **43**(4), 1–20 (2024)
115. Zhang, Y., Wang, Y., Zhang, Z., Tang, H.: Code2worlds: Empowering coding llms for 4d world generation. arXiv preprint arXiv:2602.11757 (2026)
116. Zhao, M., Nag, S., Wang, K., Vora, A., Ji, G., Chun, P., Mahdavi-Amiri, A., Zhang, H.: Advances in 4d representation: Geometry, motion, and interaction. arXiv preprint arXiv:2510.19255 (2025)
117. Zhao, Q., Zhang, X., Xu, H., Chen, Z., Xie, J., Gao, Y., Tu, Z.: Depr: Depth guided single-view scene reconstruction with instance-level diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5722–5733 (2025)
118. Zhen, H., Sun, Q., Zhang, H., Li, J., Zhou, S., Du, Y., Gan, C.: Tesseract: learning 4d embodied world models. arXiv preprint arXiv:2504.20995 (2025)

119. Zhong, C., Shi, L., Chen, H., Sun, T., Zhao, H., Yuan, B., Li, C.: Tidegs: Scalable training of over one billion 3d gaussian splatting primitives via out-of-core optimization. arXiv preprint arXiv:2605.20150 (2026)
120. Zhu, W., Li, B., Zheng, C., Mai, J., Chen, J., Jiang, L., Hamdi, A., Martinez, S.R., Lin, C.W., Elhoseiny, M., et al.: 4d-bench: Benchmarking multi-modal large language models for 4d object understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21129–21143 (2025)
121. Zhu, X., Huang, X., Xie, Q., Deng, Z., Yu, J., Guan, Y., Liu, Z., Zhu, L., Zhao, Q., Liu, L., et al.: Imaginarium: Vision-guided high-quality 3d scene layout generation. ACM Transactions on Graphics (TOG) **44**(6), 1–24 (2025)