HAT-4D: Lifting Monocular Video for 4D Multi-Object Interactions via Human-Agent Collaboration

Jiaxin Li[1,2] *, Yuxiang Wu[1] *, Zhenkai Zhang[1], Xinrui Shi[1], Haoyuan Wang[1], Yichen Zhao[1] ‡, Su Linxiang[1] ‡, Chenyang Yu[1], Mingyu Zhang[1], Yifan Ding[2], Boran Wen[1,2], Li Zhang[3], Ruiyang Liu[4], and Yong-Lu Li[1,2] †

[1] Shanghai Jiao Tong University, China [2] Shanghai Innovation Institute, China
[3] University of Science and Technology of China, China [4] Math Magic, China
{li_jiaxin, vx.limo, yonglu_li}@sjtu.edu.cn

Abstract. Extracting dynamic 4D object interactions from massive, in-the-wild monocular videos offers a highly efficient data collection pathway for scaling Embodied AI and training VLAs. However, existing monocular 4D reconstruction methods primarily focus on isolated objects, often failing under the severe occlusions and complex dynamics inherent in multi-object interactions. To bridge this gap, we propose **HAT-4D**, the first *agentic framework* designed to reconstruct the 3D geometry, temporal dynamics, and physical interactions of multiple objects from a single video. By integrating VLMs with a multi-level *human-in-the-loop* feedback mechanism, HAT-4D efficiently resolves depth ambiguities and interaction-induced occlusions during 3D generation and 4D propagation, yielding physically plausible assets without relying on expensive multi-camera rigs. As a scalable data engine, HAT-4D facilitates the creation of **MVOIK-4D**, an open-world benchmark for monocular 4D interaction reconstruction, accompanied by a novel multi-dimensional evaluation protocol focused on physical plausibility and temporal consistency. Extensive experiments demonstrate that HAT-4D achieves SOTA performance on most evaluation metrics, while maintaining competitive semantic alignment. Ablation studies show that introducing a small amount of human feedback improves interaction reconstruction. Moreover, the data produced by HAT-4D effectively improves baseline performance when used for finetuning. Our data and code are available at project webpage.

Keywords: Object Interaction · 4D generation · Agent

1 Introduction

The ability to understand and reconstruct physically plausible object interactions within a spatio-temporally coherent 4D frame is essential for numer-

* Equal contribution.
† Corresponding author.
‡ Conducted during an internship at Shanghai Jiao Tong University.

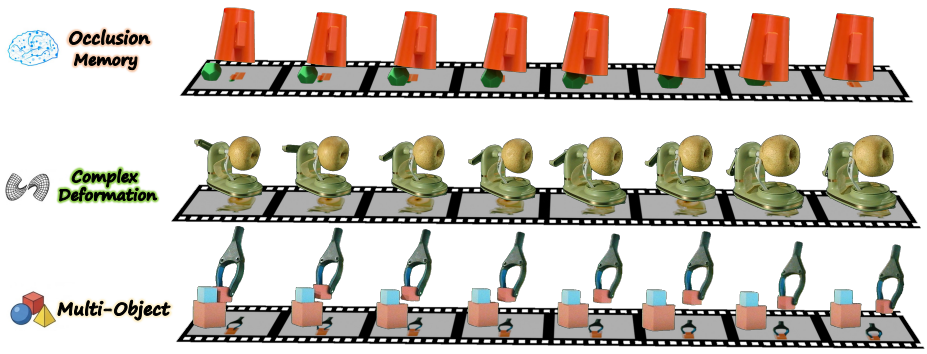


Fig. 1: Examples of object interaction sequences from the MVOIK-4D dataset. The dataset captures diverse real-world interaction scenarios with rich spatio-temporal dynamics. It includes challenging cases such as **occlusion memory**, where objects disappear and reappear across time, **complex deformation** caused by physical manipulation, and **multi-object interactions** involving coordinated motion between multiple objects. These scenarios require models to reason about geometry, dynamics, and long-term temporal consistency simultaneously.

ous downstream applications, such as populating immersive virtual environments [19, 26], and training robotic perception [9], understanding [27, 51], and manipulation [16,43,47]. Since monocular video serves as the most abundant and accessible source of real-world object interactions, lifting these 2D observations into dynamic 4D multi-object interactions has emerged as a central objective for the research [24, 33, 49]. However, this reconstruction is a fundamentally ill-posed problem, requiring the disentanglement of complex, evolving spatial relationships, the resolution of heavy mutual occlusions, and the maintenance of strict spatio-temporal consistency across the sequence.

Existing approaches to acquiring 4D object interaction data generally fall into two extremes. On the one hand, hardware-intensive approaches [8, 30] rely on massive, synchronized multi-camera rigs. While they capture occlusion-free objects with high-quality dynamics, these systems are prohibitively expensive, strictly confined to controlled studio environments, and fundamentally unscalable for open-world data collection. On the other hand, recent generative approaches attempt to bypass these hardware constraints by leveraging diffusion-based priors or synthetic data to either animating static 3D collections [11,12,18] towards 4D settings [5, 32] or to reconstruct the spatial frame from monocular video inputs [24, 31, 33, 48, 49]. However, current generative models primarily focus on isolated, stylized assets, such as gaming or cartoon characters. When confronted with the complex and diverse object interactions present in the real world, they suffer from severe domain gaps that lead to unstable generation quality, manifests as implausible physical-deformations, floating interactions, and noticeable temporal jittering.

To break the deadlock between the unscalable cost of multi-view capture and the unreliable quality of generative models, we propose **HAT-4D**, a low-cost, *human-in-the-loop (HITL)* agentic framework for monocular 4D object interaction generation. Given an in-the-wild monocular video, HAT-4D first leverages Vision-Language Model (VLM) agents to organize the visual sequence into a structured **Interaction Knowledge Graph (IKG)**. Serving as the core causality engine of our framework, the IKG efficiently encodes long-term physical changes and the underlying interaction cues that drive them. Guided by IKG, HAT-4D orchestrates a suite of specialized agents: it couples *3D object generation and composition skills* to recover the precise geometry and spatial alignment of interacting entities, and subsequently applies *4D propagation skills* to reconstruct their evolving dynamics. To explicitly resolve the severe depth and occlusion ambiguities inherent to monocular video, we integrate high-level human knowledge into the generation loop via a novel HITL collaborative scheme, empowering users to actively guide and refine the 3D generation, spatial composition, and dynamic 4D propagation processes.

Leveraging HAT-4D as a highly scalable data engine, we construct **MVOIK-4D (Multi-View Object Interaction 4D Knowledge)** benchmark encompassing 77 tasks across 112 distinct interaction scenarios, alongside a novel multi-dimensional evaluation protocol designed to rigorously assess physical plausibility and interaction stability involving deformation realism, interaction consistency, temporal smoothness, cross-view and long-term memory preservation. Extensive experiments validate that HAT-4D consistently outperforms existing monocular 4D baselines in modeling complex multi-object interactions, paving the way for downstream Embodied AI research by supplying high-fidelity, scalable physical priors.

In summary, our main contributions are: (1) We propose **HAT-4D**, an innovative human-in-the-loop multi-agent system designed to reconstruct physically plausible and temporally coherent 4D multi-object interactions directly from monocular videos. (2) We introduce **Interaction Knowledge Graph (IKG)** as the core causality engine of our framework. By explicitly encoding long-term physical changes, the IKG guides our specialized 3D generation and 4D propagation agents to resolve severe depth ambiguities and mutual occlusions. (3) We establish **MVOIK-4D**, a comprehensive benchmark encompassing a wide range of interaction scenarios, coupled with a novel multi-dimensional evaluation protocol to assess physical plausibility and temporal consistency.

2 Related Work

2.1 Object Interaction Knowledge Understanding

Early studies on object interaction understanding mainly focus on image-level reasoning, such as object relation detection [22, 25, 29] and object affordance detection. [7, 13] With the development of vision-language models (VLMs), Recent works extend this direction to long videos, enabling richer interaction understanding through temporal reasoning [10, 42].

With the progress of 3D reconstruction [15, 28, 41] and dynamic scene modeling [44, 50], object interaction understanding in the 4D domain has begun to attract attention. In 4D settings, interactions involve complex spatio-temporal dynamics beyond static semantic relationships. However, collecting large-scale real-world 4D interaction data is difficult. As a result, existing datasets are still largely limited to synthetic environments.

In this work, we study real-world 4D object interaction understanding with an agent-driven framework. We construct a multi-view object interaction dataset with rich annotations covering diverse interaction scenarios. To explicitly represent complex interactions in dynamic scenes, we introduce an **interaction knowledge graph** that models object relationships, interaction categories, and deformation states over time. This structured representation captures the underlying spatio-temporal interaction dynamics and provides guidance for reconstructing 4D object interactions from monocular videos.

2.2 Monocular Video-based 4D Generation

With the development of the video diffusion model [2, 34, 38] and a large-scale dynamic 3D object dataset [11, 12, 23, 52], 4D content generation attracts more and more attention. Recent research in 4D generation can be broadly divided into two paradigms. The first paradigm is optimization-based, leveraging pretrained image or video diffusion models through Score Distillation Sampling (SDS) to extract 4D features [14, 20, 24, 48]. However, Score Distillation Sampling (SDS) often suffers from the Janus problem when generating the complex object. The second paradigm follows a dataset-driven, end-to-end approach that relies on large-scale 3D or 4D datasets [6, 33, 45, 49]. Models such as SV4D [45], L4GM [33], and GVF-Diffusion [49] extend high-fidelity 3D generation architectures by introducing spatio-temporal cross-attention layers, enabling end-to-end dynamic scene generation. However, these models are typically trained on large-scale synthetic datasets. When applied to real-world scenes, their performance often degrades due to the significant domain gap.

To address these challenges, we propose **HAT-4D**, an agent-driven framework that integrates multi-level human-in-the-loop feedback into the 4D generation process. By introducing human corrections during generation, the framework effectively reduces error accumulation and improves the physical plausibility of generated interactions. It also mitigates memory degradation when objects undergo heavy occlusions or reappear after long temporal gaps. Furthermore, HAT-4D serves as an efficient 4D data engine to construct the **MVOIK-4D** benchmark, alleviating the lack of real-world 4D object interaction data.

2.3 Human-in-the-loop Tool in Visual Tasks

Human-in-the-loop (HITL) methods have long been closely related to data construction and the scarcity of annotated data. Early studies mainly explored HITL

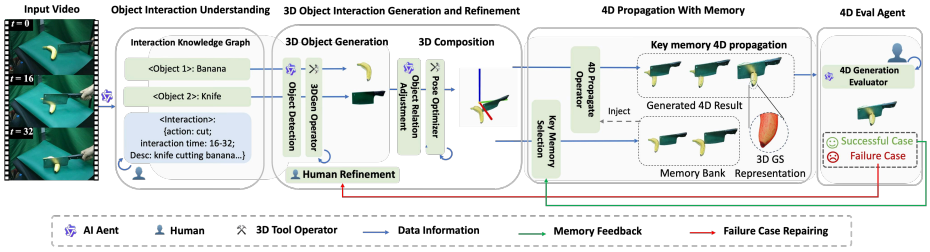


Fig. 2: Overview of HAT-4D. The framework extracts an **interaction knowledge graph** from a monocular video, reconstructs and composes the interacting objects, and generates subsequent 3D states through **memory-guided** 4D propagation. The evaluation agent updates the memory with successful results and feeds failed cases back for repair, while human users can refine intermediate outputs throughout the pipeline.

through active learning [35], where models identify uncertain samples and request human annotations, and such strategies have been widely applied in object detection [46] and medical imaging tasks [3, 40].

With the development of deep learning, HITL has evolved into a data engine paradigm. For example, Segment Anything [21] introduces a human–model data engine that iteratively expands large-scale segmentation datasets through model prediction and human correction, enabling strong segmentation performance in open-world scenarios. Recent works extend this idea to the 3D domain [4, 39], where model predictions are refined through human evaluation or expert correction to construct large-scale 3D training data and alleviate the scarcity of real-world 3D datasets.

In this work, we further extend this paradigm to dynamic 4D reconstruction. Leveraging the few-shot capabilities of vision–language models (VLMs), we propose HAT-4D, a human–agent collaborative framework that addresses the scarcity of real-world 4D interaction data. Our agent performs automatic repair during generation, while human feedback is introduced at sparse keyframes to correct accumulated errors, enabling reliable reconstruction of complex object interactions.

3 Method

Our objective is to reconstruct 4D multi-object interactions from monocular videos, a task hindered by severe depth ambiguities, heavy occlusions, and evolving topological states. To overcome these challenges and mitigate error accumulation, we propose **HAT-4D**, an agentic framework (Fig, 2) driven by an explicit **Interaction Knowledge Graph (IKG)**. Within this framework, interacting entities and their physical deformations are reconstructed as 4D Gaussian Splats [44] and organized into critical interaction event segments. Guided by the IKG, our specialized agents ensure physically plausible 3D generation and temporally consistent 4D propagation, while multi-level **human-in-the-**

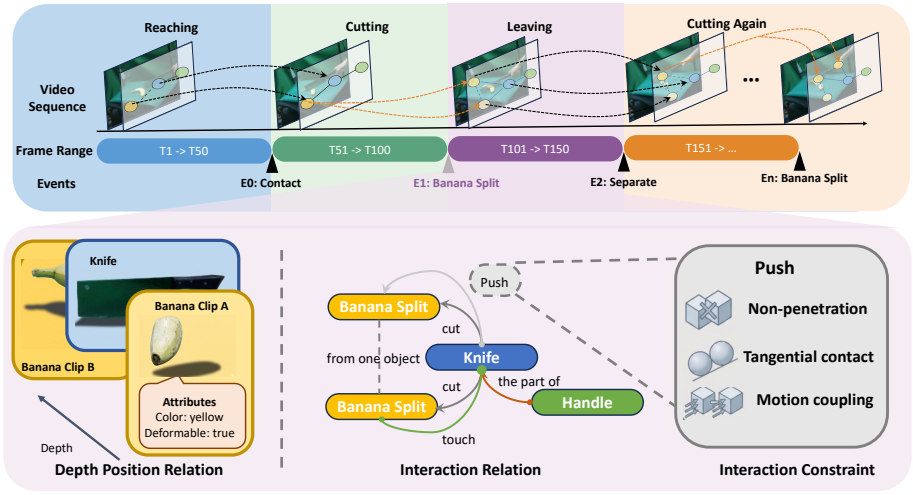


Fig. 3: Overview of the Interaction Knowledge Graph (IKG) formulation. The IKG represents physical interactions as a dynamic, temporally conditioned directed graph. **Top:** The video sequence is partitioned into event segments ($\mathcal{E}$) according to state transitions (e.g., *E0: Contact*, *E1: Banana Split*). Each segment models interaction phases such as *Reaching*, *Cutting*, and *Leaving*. **Bottom Left:** Depth cues and physical attributes (e.g., color, deformability) are extracted for object entities ($\mathcal{O}$). **Bottom Center & Right:** Spatio-temporal relations ($\mathcal{R}$) and interaction constraints (e.g., non-penetration, motion coupling) encode semantic and physical dependencies between objects (e.g., *Knife* and *Banana*), guiding downstream 4D generation.

loop (HITL) collaboration enables precise geometric refinement and semantic editing. The remainder of this section is organized as follows: Sec. 3.1 details the formulation of the **IKG**; Sec. 3.2 describes the **HAT-4D** agentic system, encompassing 3D generation, spatial composition, and memory-augmented 4D propagation skills; and Sec. 3.3 introduces our multi-level **HITL** collaborative mechanisms.

3.1 Interaction Knowledge Graph Formulation

The fundamental novelty of HAT-4D lies in leveraging the understanding of physical interactions to guide the generation of visual assets. Given an input monocular video, we first prompt Qwen3-VL [1] to digest the entire sequence and formalize it into an IKG (Fig. 3). Mathematically, the IKG is defined as a dynamic, temporally conditioned directed graph $\mathcal{G} = (\mathcal{O}, \mathcal{E}, \mathcal{R})$ that tracks object entities $\mathcal{O}$, interaction event segments $\mathcal{E}$ over the video duration T, and spatial and semantic relationships $\mathcal{R}$ between those entities within each segment.

Object Entities. $\mathcal{O} = \{o_1, o_2, ..., o_n\}$ represents the set of distinct interactable objects in the scene (e.g., banana and knife). Each object o_i encapsulates

semantic metadata and physical attributes that hints the downstream generation and validation, including colors, textures, deformability, and affordance.

Interaction Event Segments. $\mathcal{E}$. The temporal structure of the interaction is represented by a set of event segments $\mathcal{E} = \{e_1, e_2, ..., e_m\}$, partitioned by state changes, such as the severing of an object, and occlusion relation shifts. Each segment $e_i \in \mathcal{E}$ is defined by its temporal boundaries $[t_{start}, t_{end}]$ and a specific interaction phase, such as start, active, or end. These segments partition the video duration T into manageable chunks, anchoring the memory bank to ensure temporal consistency through complex interaction transitions and topological updates. In each event segment, we flag a keyframe $\ddot{T}_i \in e_i$ with maximum visual quality and minimal occlusion.

Spatio-Temporal Relations. $\mathcal{R} = \{\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_m\}$ describe the interactions occurred within each interaction event segment. Within a specific segment e_m, the relation set is defined as $\mathbf{R}_m = \{r_1, r_2, ..., r_k\}$, with each individual relation r_i defined as a tuple $r_i = \langle (o_a, o_b), \mathcal{I}, \mathcal{O}_{depth}, \mathcal{S} \rangle$, involving 4 facets:

- *Interacting Object Pair* $(o_a, o_b) \in \mathcal{O} \times \mathcal{O}$ defines the pair of entities that carries the interaction.
- *Interaction Semantics* $\mathcal{I} = \langle p, c, d \rangle$ defines the semantic predicate p of the interaction (e.g., *cutting*), accompanied by a confidence score $c \in [0, 1]$ and a distance hint d to guide spatial initialization. Specifically, the semantic predicates p represents dominant interactions between a pair of entities (o_a, o_b). We use fine-grained atomic verbs *e.g.*, $^{<o_a>}$*knife* – $^{<p>}$***cuts*** –$^{<o_b>}$*apple*.
- *Depth Ordering* $\mathcal{O}_{depth} \in \{\prec_{depth}, \succ_{depth}, \emptyset\}$ establishes the occlusion relation between entities, where $o_a \prec_{depth} o_b$ indicates that o_a consistently occludes o_b over current interaction segment.
- *Relative Position* $\mathcal{S} \in \mathcal{H} \times \mathcal{V} \times \mathcal{D}$ provides categorical alignment constraints across the horizontal (left, right, overlap), vertical (above, below, overlap), and depth (front, behind) axes.

3.2 4D Generation and Propagation with HAT-4D

With the IKG established as a formal constraint engine, HAT-4D orchestrates a multi-agent system to execute the reconstruction. The system transitions from static 3D lifting to dynamic 4D propagation by invoking 3D generation, composition, and 4D propagation skills, with automatic evaluation and rollback to mitigate error accumulation.

3D Object Generation and Composition. The generation process begins with the **3D Object Generation Skill**, which initializes the scene at both the first frame and at keyframes $\ddot{T}$ flagged by IKG. For each frame, the agent first produces a set of virtual anchors for each entity identified by the IKG and then invokes SAM3D [37] to localize and reconstruct these individual entities as 3D Gaussian Splats. Since heavy occlusions may cause multiple anchors to be assigned to a single physical entity, the agent leverages object detection to identify and consolidate such instances. For 3D objects reconstructed from multiple anchors belonging to the same physical entity, the agent selectively retains

the candidate that most closely aligns with the object semantics specified in the IKG and exhibits the highest reconstruction quality.

Following reconstruction, the **3D Object Composition Skill** integrates the independent assets into a coherent 3D scene. Guided by the relative spatial timelines defined in the IKG, the agent leverages a lightweight pose optimization operator to refine the 6DoF positions and orientations of each entity. Furthermore, the agent computes exact contact points between entities, if applicable, to ensure the resulting composition is physically plausible and interaction-consistent. Finally, these composed 3D results from the first frame and selected keyframes are cached in a **memory bank**, serving as robust spatial anchors for subsequent 4D propagation.

Memory-Augmented 4D Propagation. To extend the composed 3D scene across time, HAT-4D employs a segment-wise **4D Propagation Operator** (L4GM [33]), where we render 4 orthogonal planes from the 3D composition result as spatial initialization. The 4D propagation operator is conditioned on both immediately preceding refined frames from the first frame and keyframes in the memory bank, ensuring long-term temporal stability and maintaining strict object identity consistency throughout highly dynamic interaction phases.

Multi-Dimensional Evaluation and Rollback. To systematically prevent error accumulation, a dedicated 4D Generation Evaluation Skill serves as the system's critic. For each propagated 4D segment, the agent renders recent frames into multi-view videos and performs a comprehensive assessment driven by the IKG constraints across two dimensions:

- *Dynamic Assessment* evaluates physical plausibility (e.g., interpenetration violations) and long-term memory stability over the temporal sequence;
- *Static Assessment* examines visual fidelity, texture quality, and cross-view consistency of the individual assets.

When the agent identifies an error, it provides detailed diagnostic descriptions to the corresponding generative agents and triggers a selective rollback. Errors related to dynamic physics trigger a localized re-generation at the 4D propagation stage, whereas artifacts related to static visual quality lead to a complete rollback to the 3D object generation stage for geometric refinement.

3.3 Multi-Level Human-in-the-Loop Collaboration

Although vision–language models (VLMs) exhibit strong generalization capabilities in understanding physical knowledge under open-world settings, monocular video-based 4D generation remains a highly ill-posed problem. In ambiguous scenarios where multiple plausible interpretations exist, human knowledge is still essential. In this subsection, we describe how human knowledge is incorporated at different stages of the generation pipeline through multi-level refinement operators, guiding VLMs toward more reliable dynamic 3D generation.

Online Fine-tuning with Human Feedback. During the generation of specific video segments, when the output produced by VLM-driven agents is

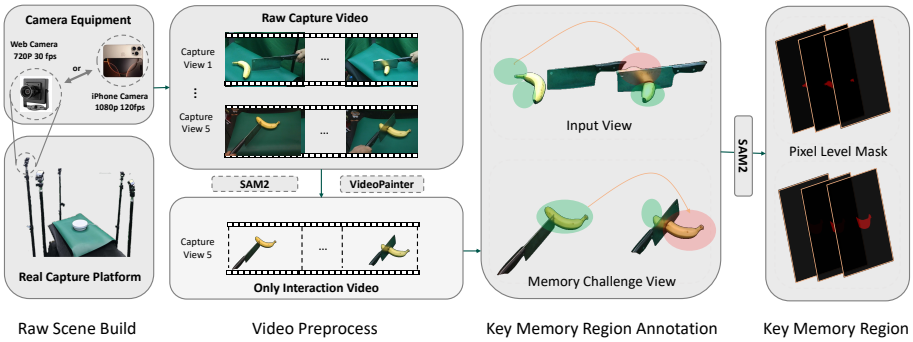


Fig. 4: Construction of MVOIK-4D and annotation of MemoryOIK-4D. We built a multi-view capture platform to record diverse 4D object interaction scenes. SAM-2 and VideoPainter are used to segment interacting objects and complete occlusions caused by hand manipulation. For memory scenarios, we select viewpoints where object visibility changes during interaction and annotate the corresponding regions with pixel-level masks to define key memory regions.

unsatisfactory, human users can directly correct the generated results. These corrections are then formulated as system prompts and provided to the corresponding agents in subsequent invocations. In this way, human knowledge is injected into an online fine-tuning agent, enabling the system to adapt its behavior with minimal additional cost. The specific prompt formats used by different agents are detailed in the appendix.

Multi-level Human Refinement Operators. To support efficient human-guided refinement of 3D generation results, HAT-4D provides a set of multi-level refinement operators.

- *Gaussian-level Refinement:* Users can directly modify attributes of selected Gaussian Splats, including position, orientation, color, and opacity.
- *Region-level Refinement:* Users can select regions with poor generation quality, where local re-generation and optimization are performed using multi-view video generation models via local latent denoising.
- *Object-level Refinement:* Specific objects can be re-generated using SAM3D, followed by pose adjustment to better align with the interaction context.

Beyond interactive editing, HAT-4D leverages these human-refined outputs to establish a continuous self-improving data engine. The generated 4D assets through the collaboration process serve as high-fidelity pseudo ground-truth for the offline fine-tuning of the underlying learnable 4D generation operators, establishing a scalable bridge for producing robust physical priors in real-world Embodied AI applications.

4 Benchmark and Metrics

To systematically evaluate the performance of our proposed HAT-4D framework in reconstructing complex, dynamic object interactions, we introduce a novel benchmark dataset **MVOIK-4D**, alongside a tailored multi-dimensional evaluation protocol involving overall visual fidelity, spatio-temporal memory retention, and the physical plausibility of the reconstructed interactions.

4.1 MVOIK-4D: Multi-View Object Interaction Knowledge Dataset

The Multi-View Object Interaction Knowledge dataset (MVOIK-4D) is constructed using animated 3D assets and a custom multi-camera capture system, with the following two complementary subsets:

- **ToolOIK-4D** focuses on tool-based interactions involving complex deformations. It covers diverse tools (e.g., knife, tongs, lighter) and targets (rigid solids, deformable solids, liquids), featuring 6 highly dynamic interaction categories like liquid splashing, fruit slicing, and peeling.
- **MemoryOIK-4D** targets spatio-temporal reasoning in scenarios with dynamic mutual occlusions. It includes *shell games*, *content-revealing interactions under viewpoint changes*, and *tool-based interactions under occlusions*, with few-shot annotations from specific input viewpoints, along with regions of interest that are critical to memory-dependent reconstruction.

As in Fig. 4, we built a multi-camera capture system to acquire real-world object interaction data with calibrated multi-view observations (see Appendix for details). For the **MemoryOIK-4D** subset specifically, we identify input viewpoints where object visibility undergoes deliberate changes across views, and leverage video segmentation annotation tools to label reconstruction regions that are closely related to memory-dependent reasoning.

4.2 Multi-Dimensional Evaluation

Standard 3D/4D generation metrics are insufficient for capturing the physical nuances of interacting entities. Therefore, we propose a multi-dimensional evaluation protocol tailored for dynamic multi-object scenarios.

Overall Generation Quality. Following Consistent4D [20], we evaluate the baseline generation quality and temporal consistency by computing CLIP, LPIPS, and FVD scores between the ground-truth multi-view videos captured by multi cameras and the corresponding rendered outputs generated by the models.

Temporal Memory Quality. We evaluate the model's temporal memory capability across three distinct levels:

- *Frame-to-Frame Continuity:* We adopt the optical-flow-aligned residual metric from VBench [17] to measure high-frequency temporal jitter between consecutive frames. We calculate the gradient difference between the ground-truth frame I_t and the warped previous frame $\hat{I}_t$, call it intra-quality. Unlike

MSE, which averages out pixel intensities, this metric is sensitive to texture flickering and fine-grained artifacts that are often perceptually disturbing but numerically small in standard metrics.
– *Long-Term Memory Stability:* To evaluate the long-term object permanence, we utilize DINOv3 [36] to split one picture into square "patches" and extract the features of every patch. Then we evaluate the similarity of two different patches by cosine-similarity. We define memory patches as "those patches similar to some patch appearing in the past of input, but not visible at present". The final score can be calculated as the ratio of the number of memory patches between the predicted picture and the gt picture. The detailed analysis of this metric is provided in the appendix.

Interaction Reconstruction Quality. We utilize Qwen3-VL [1] to qualitatively rank physical plausibility based on the realism of object deformations and the accuracy of spatial bounds during interaction (e.g., if interpenetration occurred or not). To further validate the reliability of the Qwen3-VL evaluation, we conduct a user study comparing human judgments with the model's rankings, demonstrating strong consistency between them. The detailed prompt design for Qwen3-VL and the correlation analysis with human evaluations are provided in the appendix.

5 Experiment

5.1 Setting

Baselines. To analyze the capability of existing 4D generation models in handling dynamic object interactions, we select several representative end-to-end 4D generation methods as baselines. These models include L4GM [33], GVFDiffusion [49], SV4D [45], as well as SDS-based approaches such as STAG4D [48] and FB4D [24]. We adopt the open-source Qwen-VL 235B-A22B-Instruct model as the vision–language model (VLM), SAM3D and L4GM [33] as 3D/4D generation operators respectively. We use $lr = 1 \times 10^{-5}$ for the pose optimizer, and set the memory bank size to 8, where the model generates 3D Gaussians for the subsequent 7 frames conditioned on the current frame. As ablation, HAT-4D (w.o. IKG) generates only a simple description of the interaction scene instead of a highly structured and informative IKG. HAT-4D (w.o. memory) limit the size of memory bank is zero.

5.2 Comparative Experiment

Tab. 1 reports the performance of HAT-4D, its ablated variants and other monocular video-to-4D baselines on the MVOIK-4D benchmark across multiple evaluation dimensions.

Overall quality. HAT-4D achieves the best performance on both LPIPS and FVD, indicating significantly improved perceptual quality and temporal realism

Table 1: Quantitative comparison on the MVOIK-4D benchmark. Overall generation quality, interaction reconstruction quality, and temporal memory quality are evaluated using decomposed metrics. *Ideal* indicates the reference optimal value.

Method	Generation Quality			Interaction Quality		Memory Quality	
	CLIP ↑	LPIPS ↓	FVD ↓	Deform ↑	Relation ↑	Intra ↓	Long ↑
Ideal	1.0	0.0	0.0	10	10	0.0	1.0
L4GM [33]	0.8259	0.1968	962.8593	2.5268	1.8973	0.0013	0.0598
GVFDiffusion [49]	0.8059	0.1961	1006.0087	2.5871	1.8549	0.0017	0.0494
SV4D [45]	0.7914	0.1933	1362.2214	2.4732	2.0536	0.0017	0.0412
STAG4D [48]	0.8167	0.1674	1013.8370	2.0915	1.5692	0.0011	0.0650
FB4D [24]	0.8307	0.1576	889.7666	2.3326	1.8929	0.0011	0.0468
HAT-4D (wo IKG)	**0.8370**	0.1435	**822.5620**	3.2835	2.4576	0.0006	0.0708
HAT-4D (wo memory)	0.8319	0.1467	834.1219	3.4213	2.5579	0.0006	0.0621
HAT-4D (Ours)	0.8275	**0.1433**	830.4060	**3.4487**	**2.5737**	**0.0005**	**0.0901**

of the generated 4D sequences. This improvement demonstrates the effectiveness of the agent-driven refinement pipeline. It progressively enhances generation quality through iterative interaction-aware optimization.

Interaction quality. On the interaction-specific metrics, HAT-4D shows substantial improvements in both deformation accuracy and relational consistency. This suggests that our approach better captures physically plausible object deformations while maintaining correct spatial relationships between interacting objects throughout the dynamic process.

Memory consistency. HAT-4D achieves state-of-the-art performance across all temporal memory metrics. In particular, our method significantly improves both short-term temporal smoothness and long-horizon consistency. This result demonstrates strong capability in preserving coherent object states across extended temporal sequences and multiple viewpoints.

Key Module Ablation. Adding IKG improves interaction reconstruction and memory consistency. It increases the Deform, Relation, Intra, and Long scores. However, the longer context from the structured IKG slightly reduces CLIP and FVD performance, likely because it weakens VLM-based discrimination. The memory module mainly improves long-term consistency. It preserves object states and maintains temporal stability over long sequences.

Qualitative Analysis. As shown in Fig. 5, we present qualitative comparisons between HAT-4D and several baseline methods for monocular reconstruction of object interactions. The results demonstrate that the progressive evaluation and generation strategy in HAT-4D produces more stable dynamic reconstructions during the generation process.

For example, in the clip–card interaction, the shape of the clip remains clear and consistent across frames, while other methods suffer from structural instability. Moreover, HAT-4D reconstructs interaction regions more accurately. In challenging cases, such as knife cutting a banana and lighter ignition, our method preserves sharper geometry and clearer interaction boundaries.

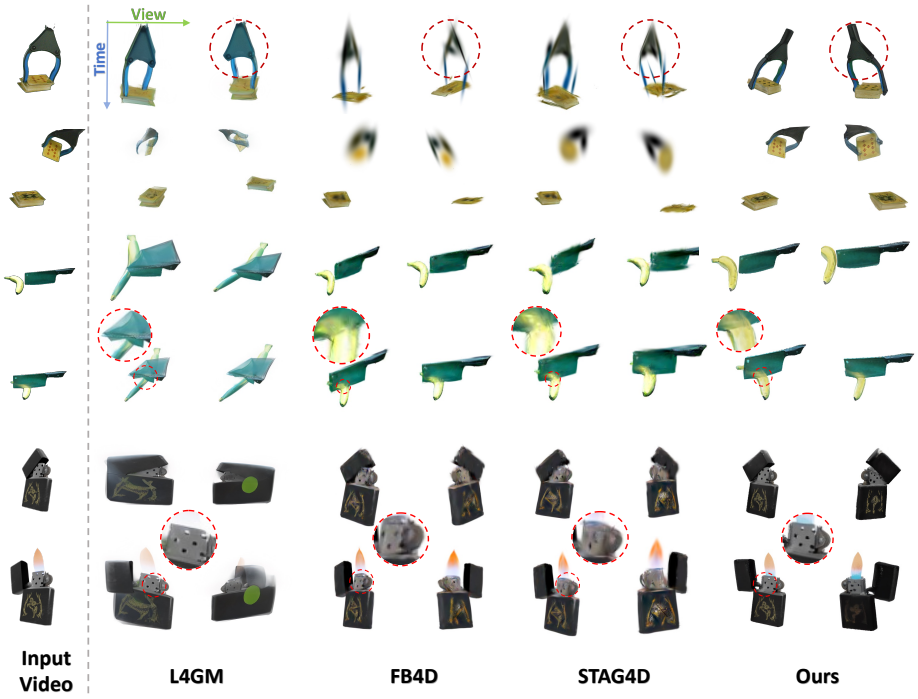


Fig. 5: Qualitative comparison with baseline methods. Compared with existing approaches, the agent-driven **HAT-4D** produces clearer 3D structures and more stable reconstructions in object interaction regions, resulting in more consistent dynamic object interaction.

In addition, HAT-4D better captures changes in the manipulated objects. For instance, the reconstructed flame exhibits clearer structure and more stable temporal dynamics compared with baseline methods.

5.3 Ablations about Human Intervention

Different Human Intervention. We study how different levels of human intervention influence the performance of HAT-4D. The experiment is conducted on 39 sequences from the MVOIK-4D dataset, among which 17 sequences contain challenging memory-intensive interactions. During generation, we limit the maximum number of human annotations and evaluate the performance under different human intervention frequencies.

As shown in Tab. 2a and Fig. 6a, even limited human intervention consistently improves reconstruction quality, interaction accuracy, and temporal consistency over the fully agent-driven setting.

The improvement is particularly significant when only a few interventions are allowed. Correcting several critical frames is sufficient to fix major errors in

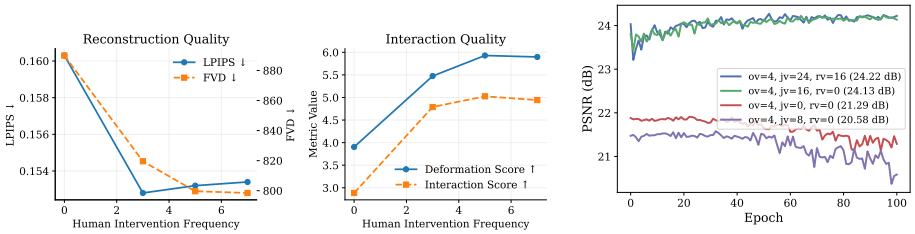
Table 2: Ablation studies on human intervention and refinement operators.

(a) Human-intervention budget

HIFs	Generation Quality			Interaction		Memory	
	CLIP↑	LPIPS↓	FVD↓	Deform↑	Relation↑	Intra↓	Long↑
0	0.8341	0.1603	889.9583	3.9038	2.8846	0.0005	0.0012
3	0.8489	**0.1528**	819.5809	5.4744	4.7885	0.0005	0.0026
5	**0.8521**	0.1532	799.6682	**5.9295**	**5.0256**	0.0005	**0.0027**
7	0.8518	0.1534	**798.4898**	5.8974	4.9423	0.0006	0.0023

(b) Multi-level refinement operators

Setting	Generation Quality			Interaction		Memory	
	CLIP↑	LPIPS↓	FVD↓	Deform↑	Relation↑	Intra↓	Long↑
Agent	0.8330	0.2126	860.5518	4.1000	3.0250	0.0006	0.0013
Obj.	0.8641	0.2060	685.7023	**6.1750**	5.1500	0.0006	0.0014
Obj.+Reg.	0.8657	0.2051	**681.7343**	5.8000	5.1500	0.0006	**0.0016**
Obj.+GS	**0.8659**	**0.2044**	686.9764	5.8750	**5.2500**	0.0006	0.0014



(a) Impact of human intervention on L4GM finetuning. HIF denotes the number of human annotations allowed during generation, with HIF= 0 representing a fully agent-driven pipeline. A small number of human interventions substantially improves reconstruction and interaction quality.

(b) View scaling. L4GM finetuning under different input-view configurations on HAT-4D sequences. *ov*, *jv*, and *rv* denote original, jittered, and randomly sampled views, respectively.

Fig. 6: Impact of human intervention and view scaling on 4D reconstruction.

the generated dynamics and interaction states. As the number of interventions increases further, the performance gain gradually saturates. This observation indicates that sparse human knowledge can effectively guide the generation process and substantially improve the reliability of dynamic 4D reconstruction.

Different Refinement Operators. We evaluate different refinement operators on 10 randomly sampled cases. For each setting, volunteers can only use the specified refinement tools. The fully agent-driven pipeline serves as the baseline.

As shown in Tab. 2b, object-level refinement yields the largest gain by correcting geometry, pose, occlusion, and interaction errors. Region-level and Gaussian-level refinements provide complementary improvements: the former achieves the best FVD and Long scores, while the latter performs best on CLIP, LPIPS, and Relation by reducing local artifacts and appearance inconsistencies.

Overall, object-level refinement is the primary correction tool. Region-level and Gaussian-level operators are used for fine-grained local adjustments. This is consistent with our practical annotation workflow.

Baseline Finetuned in Different Views We select 63 high-quality annotated scenes from MVOIK-4D and split them into training and test sets with a ratio of 9:1. To follow the training protocol of L4GM, each scene is divided into clips with 8 frames. This produces 1,275 training clips and 97 test clips. For each scene, we choose one real captured viewpoint as the reference view. Four orthogonal

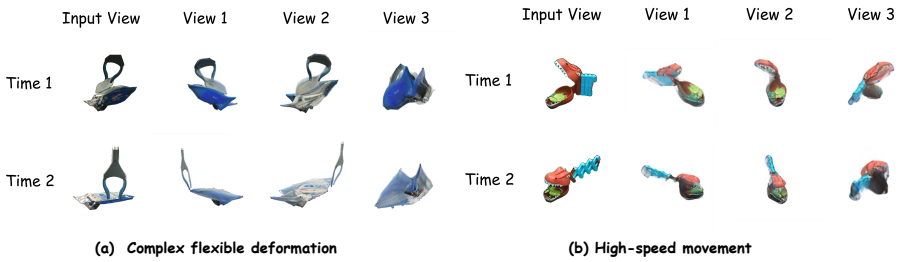


Fig. 7: Failure cases. Left: pressing and folding a thin plastic carton causes complex deformation, distorting novel-view geometry and over-smoothing details. Right: the spring toy's rapid extension and recoil cause motion blur and geometric inconsistencies.

views are rendered around it and used as input views. Another four captured viewpoints are reserved as validation views for evaluation.

To study the effect of view supervision, we render 24 jittered views from small angular and radial perturbations and 16 uniformly sampled random views. Fig. 6 shows the PSNR curves during the finetuning of L4GM on MVOIK-4D. With limited supervision views, jittered views alone do not improve the reconstruction quality of L4GM, as the model tends to overfit to a narrow set of viewpoints, which leads to unstable training and lower PSNR.

Increasing the number of supervision views improves the performance of L4GM. Randomly sampled views provide stronger geometric diversity and better spatial coverage, enabling more stable optimization and higher reconstruction accuracy. This highlights the importance of stronger multi-view supervision and the meaning of HAT-4D for dynamic 4D generation.

6 Conclusion

We present **HAT-4D**, an agentic framework for reconstructing dynamic 4D object interactions from monocular videos. By integrating multi-level human-in-the-loop feedback, HAT-4D refines 3D object generation, spatial composition, and 4D dynamic propagation, improving physical plausibility and temporal consistency. We also introduce **MVOIK-4D**, a benchmark with **112 scenes**, **77 tasks**, **39 interaction categories**, and **15 object deformation categories**, together with a multi-dimensional evaluation protocol for generation quality, interaction consistency, and long-term memory stability.

However, as shown in Fig. 7, HAT-4D remains challenged by complex flexible deformation and fast non-rigid motion, where irregular geometry and motion blur degrade SAM3D-based geometry reconstruction. Performance also depends on the reasoning capability and inference efficiency of the underlying VLM. Future work will explore improved temporal correspondence, deformation modeling, VLM fine-tuning, and more scalable 4D interaction data generation from monocular videos. We hope HAT-4D and MVOIK-4D will support research in object interaction understanding, 4D generation, and robotic perception.

Acknowledgements

This work was supported by the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G14), National Natural Science Foundation of China (U25A20442, 62306175), Ant Group.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis **71**, 102062 (Jul 2021). https://doi.org/10.1016/j.media.2021.102062, http://dx.doi.org/10.1016/j.media.2021.102062
4. Cen, J., Fang, J., Zhou, Z., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment anything in 3d with radiance fields (2024), https://arxiv.org/abs/2304.12308
5. Chen, H., Chen, X., Zhang, Y., Xu, Z., Chen, A.: Motion 3-to-4: 3d motion reconstruction for 4d synthesis. arXiv preprint arXiv:2601.14253 (2026)
6. Chen, J., Zhang, B., Tang, X., Wonka, P.: V2m4: 4d mesh animation reconstruction from a single monocular video. arXiv preprint arXiv:2503.09631 (2025)
7. Chen, J., Gao, D., Lin, K.Q., Shou, M.Z.: Affordance grounding from demonstration video to target image (2023), https://arxiv.org/abs/2303.14644
8. Chen, Y., Guo, S., Yang, T., Ding, L., Yu, X., Gu, J., Xue, T.: 4dslomo: 4d reconstruction for high speed scene with asynchronous capture. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
9. Chen, Z., Li, J., Liang, J., Tan, L., Guo, Y., Lu, C., Li, Y.L.: Mˆ 3-vos: Multi-phase, multi-transition, and multi-scenery video object segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29193–29202 (2025)
10. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., Bing, L.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms (2024), https://arxiv.org/abs/2406.07476
11. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)

12. 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)
13. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection (2018), https://arxiv.org/abs/1709.07326
14. Geng, C., Zhang, Y., Wu, S., Wu, J.: Birth and death of a rose. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26102–26113 (2025)
15. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d (2024), https://arxiv.org/abs/2311.04400
16. Hou, C., Ze, Y., Fu, Y., Gao, Z., Hu, S., Yu, Y., Zhang, S., Xu, H.: 4d visual pre-training for robot learning (2025), https://arxiv.org/abs/2508.17230
17. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models (2023), https://arxiv.org/abs/2311.17982
18. Hunyuan3D, T., :, Zhang, B., Guo, C., Guo, D., Liu, H., Yan, H., Shi, H., Yu, J., Xu, J., Huang, J., Li, K., Wang, L., Linus, Wang, P., Lin, Q., Tang, R., Yang, X., Li, Y., Guan, Y., Zhao, Y., Yang, Y., Lai, Z., Liang, Z., Zhao, Z.: Hy3d-bench: Generation of 3d assets (2026), https://arxiv.org/abs/2602.03907
19. Ji, K., Shi, Y., Jin, Z., Chen, K., Xu, L., Ma, Y., Yu, J., Wang, J.: Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis (2025), https://arxiv.org/abs/2508.02106
20. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848 (2023)
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
22. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016), https://arxiv.org/abs/1602.07332
23. Li, B., Zheng, C., Zhu, W., Mai, J., Zhang, B., Wonka, P., Ghanem, B.: Vivid-zoo: Multi-view video generation with diffusion model. Advances in Neural Information Processing Systems **37**, 62189–62222 (2024)
24. Li, J., Gao, H.a., Li, W., Chi, H., Liu, C., Du, C., Liu, Y., Gao, M., Zhang, G., Zhang, Z., et al.: Fb-4d: Spatial-temporal coherent dynamic 3d content generation with feature banks. arXiv preprint arXiv:2503.20784 (2025)
25. Li, R., Zhang, S., He, X.: Sgtr: End-to-end scene graph generation with transformer (2022), https://arxiv.org/abs/2112.12970
26. Li, S., Zhang, H., Chen, X., Wang, Y., Ban, Y.: Genhoi: Generalizing text-driven 4d human-object interaction synthesis for unseen objects (2025), https://arxiv.org/abs/2506.15483
27. Li, Y.L., Xu, Y., Xu, X., Mao, X., Yao, Y., Liu, S., Lu, C.: Beyond object recognition: A new benchmark towards object concept learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20029–20040 (2023)

28. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023), https://arxiv.org/abs/2303.11328
29. Liu, X., Wen, B., Liu, X., Zhou, Z., Fan, H., Lu, C., Ma, L., Chen, Y., Li, Y.L.: Interacted object grounding in spatio-temporal human-object interactions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 5622–5630 (2025)
30. Lu, C.Y., Zhou, P., Xing, A., Pokhariya, C., Dey, A., Shah, I.N., Mavidipalli, R., Hu, D., Comport, A.I., Chen, K., et al.: Diva-360: The dynamic visual dataset for immersive neural fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22466–22476 (2024)
31. Ma, Z., Chen, X., Yu, S., Bi, S., Zhang, K., Ziwen, C., Xu, S., Yang, J., Xu, Z., Sunkavalli, K., et al.: 4d-lrm: Large space-time reconstruction model from and to any view at any time. arXiv preprint arXiv:2506.18890 (2025)
32. Mou, L., Lei, J., Wang, C., Liu, L., Daniilidis, K.: Dimo: Diverse 3d motion generation for arbitrary objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14357–14368 (2025)
33. 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)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
35. Settles, B.: Active learning literature survey (2009)
36. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
37. 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
38. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024)
39. Wang, B., Wu, V., Wu, B., Keutzer, K.: Latte: Accelerating lidar point cloud annotation via sensor fusion, one-click annotation, and tracking (2019), https://arxiv.org/abs/1904.09085
40. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., Vercauteren, T.: Deepigeos: A deep interactive geodesic framework for medical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(7), 1559–1572 (Jul 2019). https://doi.org/10.1109/tpami.2018.2840695, http://dx.doi.org/10.1109/TPAMI.2018.2840695
41. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer (2025), https://arxiv.org/abs/2503.11651
42. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Yan, Z., Zheng, R., Xu, J., Wang, Z., Shi, Y., Jiang, T., Li, S., Zhang, H., Huang, Y., Qiao, Y., Wang, Y., Wang, L.: Internvideo2: Scaling foundation models for multimodal video understanding (2024), https://arxiv.org/abs/2403.15377

43. Wang, Z., Liu, C., Xiang, Y., Zhang, R., Hao, Q., Lu, H., Chen, H., Feng, Z., Zheng, K., Ye, D., et al.: The great march 100: 100 detail-oriented tasks for evaluating embodied ai agents. arXiv preprint arXiv:2601.11421 (2026)
44. 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 (2024), https://arxiv.org/abs/2310.08528
45. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024)
46. Yu, W., Zhu, S., Yang, T., Chen, C.: Consistency-based active learning for object detection (2022), https://arxiv.org/abs/2103.10374
47. Zeng, J., Bu, Q., Wang, B., Xia, W., Chen, L., Dong, H., Song, H., Wang, D., Hu, D., Luo, P., Cui, H., Zhao, B., Li, X., Qiao, Y., Li, H.: Learning manipulation by predicting interaction (2024), https://arxiv.org/abs/2406.00439
48. Zeng, Y., Jiang, Y., Zhu, S., Lu, Y., Lin, Y., Zhu, H., Hu, W., Cao, X., Yao, Y.: Stag4d: Spatial-temporal anchored generative 4d gaussians. In: European Conference on Computer Vision. pp. 163–179. Springer (2024)
49. Zhang, B., Xu, S., Wang, C., Yang, J., Zhao, F., Chen, D., Guo, B.: Gaussian variation field diffusion for high-fidelity video-to-4d synthesis. arXiv preprint arXiv:2507.23785 (2025)
50. Zhang, C., Moing, G.L., Koppula, S., Rocco, I., Momeni, L., Xie, J., Sun, S., Sukthankar, R., Barral, J.K., Hadsell, R., Ghahramani, Z., Zisserman, A., Zhang, J., Sajjadi, M.S.M.: Efficiently reconstructing dynamic scenes one d4rt at a time (2025), https://arxiv.org/abs/2512.08924
51. Zhang, M., Zhuo, L., Tan, T., Xie, G., Nie, X., Li, Y., Zhao, R., He, Z., Wang, Z., Cai, J., et al.: Ipr-1: Interactive physical reasoner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 33415–33425 (2026)
52. Zhang, Y., Zhang, L., Ma, R., Cao, N.: Texverse: A universe of 3d objects with high-resolution textures. arXiv preprint arXiv:2508.10868 (2025)