ByteLOOM: Weaving Geometry-Consistent Human-Object Interactions Videos through Progressive Curriculum Learning

Bangya Liu[1,†], Xinyu Gong[2], Zelin Zhao[3,†], Ziyang Song[4,†], and Suman Banerjee[1]

[1] University of Wisconsin-Madison
[2] ByteDance
[3] Georgia Institute of Technology
[4] The Hong Kong Polytechnic University
[†]Project during internship at ByteDance

Abstract. Human-object interaction (HOI) video generation has garnered increasing attention due to its promising applications in digital humans, e-commerce, advertising, and robotics imitation learning. However, existing methods face two critical limitations: (1) a lack of effective mechanisms to inject multi-view information of the object into the model, leading to poor cross-view consistency, and (2) heavy reliance on fine-grained annotations including hand mesh and body templates for modeling interaction occlusions, due to limited training data. To address these challenges, we introduce ByteLOOM, a Diffusion Transformer (DiT)-based framework that generates realistic human-object interaction video with geometrically consistent object illustration, using simplified human conditioning and 3D object inputs. We first propose an RCM-cache mechanism that leverages Relative Coordinate Maps (RCM) as a universal representation to maintain object's geometry consistency and precisely control 6-DoF object transformations in the meantime. To compensate HOI dataset scarcity and leverage existing datasets, we further design a training curriculum that enhances model capabilities in a progressive style and relaxes the demand of hand mesh. Extensive experiments demonstrate that our method faithfully preserves human identity and the object's multi-view geometry, while maintaining smooth motion and object manipulation. Project page: neutrinoliu.github.io/byteloom/

1 Introduction

The synthesis of realistic human-object interaction (HOI) videos has emerged as a critical research area with applications spanning digital commerce, entertainment, robotics, and virtual production. While recent advances in video generation have demonstrated remarkable capabilities in synthesizing human motion and appearance [32, 43, 48], generating videos featuring complex object manipulation with geometric consistency remains a significant challenge.

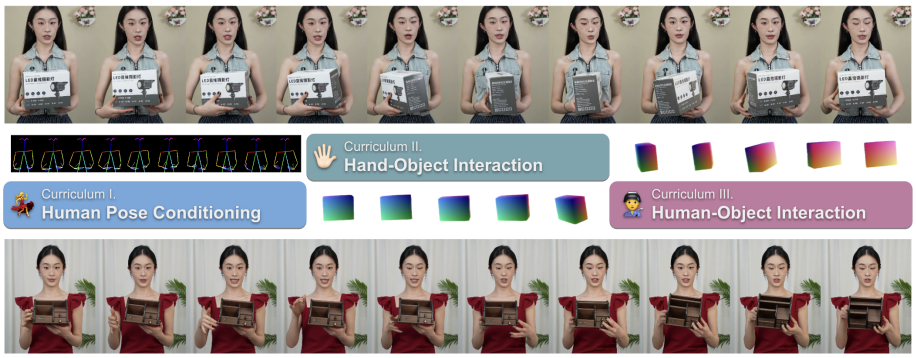


Fig. 1: ByteLOOM generates high-quality human-object interaction videos while maintaining consistent object geometry and appearance across frames. Our method introduces two key innovations: First, we propose Relative Coordinate Maps (RCM) as a universal representation for 3D object poses. Second, we develop a training curriculum that enables the diffusion model to learn HOI video generation progressively.

Recent efforts have made progress on specific aspects of this challenge. AnchorCrafter [36] as the pioneer work introduces a dedicated bundle of conditions, including human reference, human pose, hand mesh, object reference and object depth map, to achieve the HOI rendering. HunyuanVideo-HOMA [20] further relaxes the dependency on curated motion data through weakly conditioned multimodal guidance. DreamActor-H1 [29] employs masked cross-attention and a large 3D motion templates for human-product demonstration videos. Despite these advances, two fundamental limitations persist: (1) the lack of effective mechanisms to inject multi-view information into diffusion models, leading to poor geometric consistency of manipulated objects, and (2) a heavy reliance on fine-grained annotations like hand mesh that are costly to scale up.

To address these challenges, we introduce **ByteLOOM**, a DiT-based framework that generates high-fidelity HOI videos (specifically, 3rd-person product showcase video) with geometrically consistent object manipulation using simplified human conditioning and 3D object inputs.

The first innovation of our approach is a three-stage curriculum learning strategy that gradually builds the model's capability from basic human motion understanding to full human-object interaction. Because the availability of training data varies dramatically across interaction complexity, we structure the curriculum to exploit this imbalance.

By progressing from widely available pose-only data, to plentiful hand-object data, and finally to scarce full-interaction data, the curriculum equips the model with strong foundational motion understanding before introducing increasingly complex manipulation tasks. This staged design leads to better generalization, more stable learning, and significantly more consistent HOI synthesis.

Our second innovation is the RCM (Relative Coordinate Map) cache, designed to provide the diffusion model with a persistent and reliable 3D prior. Instead of relying on single-frame or one-time conditioning, we bind multi-view

reference RGB images with their corresponding RCMs and store them as a cache accessible to the DiT throughout generation. During synthesis, the model also receives per-frame RCM inputs that specify the object's target 6-DoF pose and geometry for that timestep. By jointly leveraging the cached multi-view RCM-RGB pairs as a global appearance prior and the per-frame RCM as a geometric control signal, the model implicitly learns to retrieve the correct object appearance from the cache and render it faithfully in every frame. This mechanism ensures that the object remains geometrically consistent, correctly textured, and view-aligned even under complex manipulation.

Extensive experiments demonstrate that our method preserves human identity, maintains smooth motion, and substantially surpasses prior work in object manipulation fidelity and multi-view consistency. Overall, our contribution could be summarized as follows:

- We design a structured training curriculum (comprising human-pose pretraining, hand-object pretraining, and HOI finetuning) that effectively balances the scarcity of high-quality HOI data with the progressively increasing granularity of interaction conditioning.
- We introduce RCM-cache, a universal mechanism for injecting multi-view 3D object information and precisely controlling object 6-DoF transformations. This persistent 3D prior enables substantially improved geometric consistency across video frames.
- We designed and practiced a depth-free data curation pipeline that robustly extracts HOI conditionings including human pose and object 6DoF pose.
- We establish **Mani4D**, a dedicated benchmark for evaluating geometry-consistent object manipulation in generated HOI videos. It is open-sourced at huggingface.co/byteloom-HOI.

2 Related Work

2.1 Human-Object Interaction Video Generation

Human-object interaction (HOI) generation has emerged as a critical challenge in creating physically plausible visual content.

In the **motion synthesis** domain, extensive research has focused on predicting human motions during object interactions, with methods like InterDiff [35], TeSMo [40], ROG [37], and HOIAnimator [25] generating interaction sequences based on 3D object representations. Those methods show potential of providing precise human and object pose conditioning for pixel level interaction generation.

For **image generation**, InteractDiffusion [15] provides interaction controllability through text-to-image diffusion models, while VirtualModel [5] synthesizes HOI images by leveraging input objects and human poses. HOGAN [16] addresses hand-object interactions via split-and-combine approaches considering inter-occlusion patterns. Other generalized methods such as [47] could also enhance the quality of the generation.

Within **video generation**, two primary paradigms have emerged: **(a) video editing** and **(b) generation from scratch**. Editing-focused methods like HOI-Swap [38] and ReHoLD [8] enable object replacement in existing hand-centric videos while preserving interaction realism, whereas MIMO [22] and AnimateAnyone 2 [17] substitute human subjects in complex dynamic scenes. However, these approaches remain constrained to editing existing content and often introduce visual artifacts when handling complex interactions occupying small pixel regions. For generation from scratch, AnchorCrafter [36] integrates HOI into pose-guided frameworks using multi-conditional inputs such as pose and depth, while ManiVideo [23] achieves precise control through 3D modeling of both human hands and objects. DreamActor-H1 [29] extends this route with a large library consisting numerous mesh-based templates. Nevertheless, these methods typically rely on strong conditional inputs or lack sufficient generative freedom, highlighting the need for approaches that balance controllability with generalizability across diverse objects and interaction scenarios while maintaining high-quality synthesis of full-body human-object interactions. HunyuanVideo-HOMA [20], introducing a totally different paradigm, adopts weak condition to relax this tension yet fails to display diverse viewpoints of the object.

2.2 Geometry-Consistent Generation

Geometry-consistent generation aims to preserve the underlying 3D structure, shape, and texture of objects across viewpoints and over time. A first line of work focuses on multi-view image generation as a strong 3D prior [14, 24, 45]. MVDream [24] trains a multi-view diffusion model on both 2D images and 3D data to produce sets of images that are coherent across viewpoints, combining the diversity of 2D diffusion with the cross-view consistency of 3D renderings. LucidFusion [14] similarly targets multi-view coherence by converting each input image into a Relative Coordinate Gaussian representation and using differentiable 3D Gaussian rendering to enforce globally aligned geometry without requiring camera poses. More recently, several works extend such **3D-aware priors** to videos or directly model geometry-consistent video generation. Zhang *et al.* propose 4Diffusion [42], which augments a 3D-aware image diffusion backbone with temporal attention layers to synthesize view-consistent videos. Other approaches incorporate explicit 3D control signals into diffusion models to tie video frames to stable underlying geometry. Diffusion-as-Shader [11] conditions frame synthesis on tracked 3D point trajectories and uses color anchors to bind RGB appearance to persistent 3D points, leading to stable surface textures over time. Related methods leverage object depth maps or pseudo-3D cues [26, 30, 36, 46] to guide object placement, improving multi-view and temporal consistency while still relying on 2D video backbones.

3 Method

3.1 Overview

Following pioneered works [20, 29, 36, 46], **ByteLOOM** takes inputs condition of both human and objects and generates realistic human-object interaction video

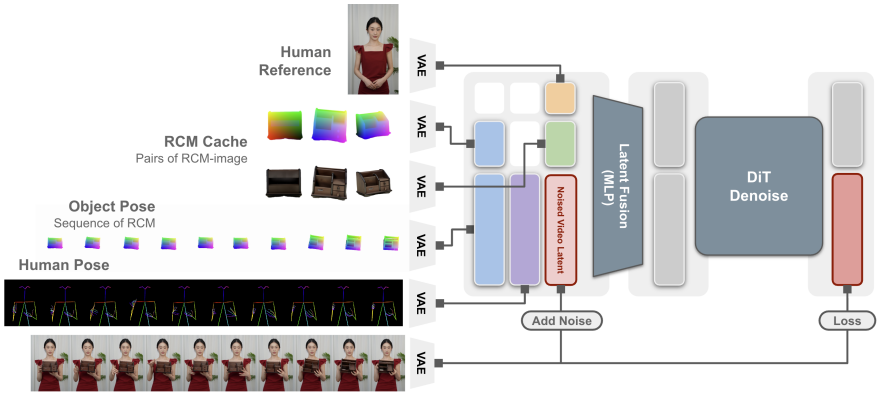


Fig. 2: Overview of **ByteLOOM** training strategy. Different kinds of conditions are concatenated into input latent. Then pass through a MLP fuser.

as outputs. For the conditioning injection, inspired by the insights from [46, 48], we plainly concat and inject all conditions through the input latent fusion, where we let the self-attention to learn the relationship between those input conditions and the final output, as long as we have provided all required information. Fig. 2 shows the overview of our system.

Compared with previous works, **ByteLOOM** distinguishes itself from two perspectives: (1) instead of a front-view dominant object presentation, we aim to present a much wider range of views for the object. This raises the requirement of an effective mechanism to deliver such multi-view information into the generation process. (2) instead of relying on precise hand mesh as the interaction conditioning [29, 36], we merely use the simple OpenPose [2] style skeleton to condition the human pose. Yet this makes it difficult for the model to directly infer the hand-object occlusion relationship via stick figure.

To tackle with the first challenge, we innovatively introduce RCM-cache module, detailedly introduced in Sec. 3.2. To better teach the model how to draw realistic interaction frames, we developed a three stage training curriculum, which will be detailed introduced in Sec. 3.3. Further, we propose a new metric in Sec. 3.4 to quantify the geometry consistency of the presented object in the generated video. Finally, we share our experience when we curate our own HOI datasets in Sec. 3.5.

3.2 Universal Object 6DoF Conditioning

Previous video generation tasks, either subject-to-video [21] or human-object interaction video [20, 29], dominantly apply single image as the object conditioning. AnchorCrafter [36], though incorporate three images as object conditions, yet all of them are within a very narrow view cone within approximately 30 degrees, which still dramatically constrains the view angles of the object in the generated video. This conditioning paradigm works for simple interactions like plainly moving the object from left to right. Yet for more complex interactions

(a) (b)

Fig. 3: (a) Model still struggles for generating geometry consistent object, even when multiview object references are provided. (b) RCM-cache of **ByteLOOM**. Sparse views are provided to the model as a reference when performing novel view synthesis.

like rotation, inverting, and manipulating, they fail the task inevitably. One of the key reasons is the lacking of multiview information, in which case model could only guess how to draw the unknown face.

A naive remedy for this problem will be providing multi-view images as a more comprehensive reference. Yet as shown in Fig. 3a, geometry shape of the objects gradually gets twisted and collapse throughout frames, especially when the multi-view images are provided in a random sequence. This could be attributes to that the model does not have the concepts of geometry consistency of a single object, instead it tackles the multiple reference images as multiple independent objects and trying to interpolate them across frames to fulfill the reference conditioning. The remedy here is to construct a strong association in between the multi-view reference images so that the model realize that those reference frames belongs to the same object.

To achieve this, we innovatively introduce relative coordinate map (RCM) into the object conditioning. RCM is originally used in [14] as an auxiliary task for 3D assets generation, which helps to enforce the correspondence between the predicted Gaussian parameters across different views. In **ByteLOOM**, we adopt it as our condition inputs as an explicit 3D prior injection.

As shown in Fig. 3b, to render a relative coordinate map (RCM), geometric surface information is encoded by normalizing vertex positions relative to the object's spatial extents. Specifically, given a vertex position $V_i \in \mathbb{R}^3$ and the object's axis-aligned bounding box defined by two corners $b_{\min}$ and $b_{\max}$ in any body diagonal, the normalized coordinate vector C_i^{RCM} is calculated via component-wise division:

$$C_i^{\text{RCM}} = \frac{V_i - b_{\min}}{b_{\max} - b_{\min}} \in [0, 1.0]^3 \tag{1}$$

$$c_i = \text{round}(255 \times C_i^{\text{RCM}}) \in [0, 255]^3 \tag{2}$$

During the rasterization stage, the rasterizer linearly interpolates the colors defined at the triangle vertices across the surface. For a specific pixel point in the triangle, the final color c_{pixel} is determined by the barycentric weights α, β, γ

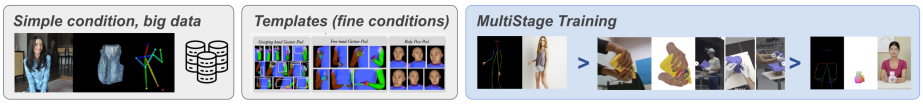


Fig. 4: Difference between the training paradigm among HunyuanVideo-HOMA [20] (simple condition + data scaling up), DreamActor-H1 [29] (fine-grained conditioning), and ours (multistage training).

of the triangle:

$$c_{\text{pixel}} = \alpha c_i + \beta c_j + \gamma c_k \tag{3}$$

This ensures that every pixel in the output map represents the exact spatial coordinate of the visible surface at that point.

When it comes to video generation, we first construct a RCM-Cache, as shown in Fig. 3b, which is a bunch of RCM and textured images pairs rendered from sparse view angles. We concat this pairs into the frame dimension of video latent as the object reference. Further, for each frame, we render out the desired RCM images with given 6DoF pose, then concat this per frame object pose condition into the channel dimension as shown in Fig. 2. In this case, we reduce the tough novel view synthesis (NVS) task into a relatively easy texturing task.

One desired property of RCM is such explicitly correspondence in between cross view pixels, so that model is always aware of pixel correlation across frames. Another good property of RCM based object condition is its universality that any object mesh with any texture could be uniformly reduced to a texture look up problem. Even applicable to articulable object as long as there is a consistent mapping between the 3D vertex and its texturing.

3.3 Staged Curriculum Learning

We have observed two distinct trends in the previous hand conditioning as shown in Fig. 4: exampled by AnchorCrafter [36] and DreamActor-H1 [29] who heavily rely on strong template based conditioning, and a totally different method exampled by HunyuanVideo-HOMA [20] who trying to achieve a scaling up with weak human pose conditioning. Both tracks actually reflect the same challenge for the HOI video generation, that there are very limited open HOI dataset with comprehensive object and human annotation.

Realizing this hard tension, we developed a third route that integrates training data from other computer vision tasks like hand-object interaction and object 6DoF estimation. Through a three stage curriculum learning procedure, we let the model learn how to animate human, how to draw geometry consistent object, and how to draw smooth and natural hand occlusion step by step.

The teaser Fig. 1 shows the curriculum I, II, III that we let the model go through in order to pick up different conditioning capabilities:

– **Curriculum I**: focuses on human-pose-conditioned training, using abundant and easily accessible human motion datasets. This stage teaches the model to generate natural human body movement, temporal dynamics, and coarse spatial structure without involving object interactions.

- **Curriculum II**: leverages the relatively rich hand-object interaction datasets, enabling the model to learn fine-grained skills such as hand-object contact reasoning, grasp configurations, and precise manipulation signals.
- **Curriculum III**: built on the much rarer full human-object interaction datasets, introduces whole-body coordination, global motion patterns, and long-range spatiotemporal consistency under object manipulation.

3.4 Quantify Geometry Consistency

We also need a metric to quantify the object geometry consistency of the generated videos. Existing metrics like MEt3R [1] only operate on image pairs and incur $O(N^2)$ computation complexity. Rolling frame-by-frame pairing could be a feasible but not perfect practice in this case.

To this end, we propose T-SSIM, a metric that leverages 3D Gaussian Splatting (3DGS) as reconstruction primitive for robust geometry consistency quantification. Specifically, we first crop out object segments from the generated video using SAM2 with uniform white padding. Given n frames $I_0, \ldots, I_n$, E-Rayzer [44] performs feed-forward 3DGS reconstruction, jointly predicting a set of 3D Gaussians $\mathcal{G}$ and camera parameters including a shared intrinsics K and per-frame extrinsics E_i:

$$\mathcal{G}, K, \{E_0, \ldots, E_n\} = \text{E-Rayzer}(I_0, \ldots, I_n) \tag{4}$$

The predicted 3DGS integrates multi-view geometry into a unified, continuous representation. We then render the 3D Gaussians back to each frame's viewpoint via differentiable splatting:

$$I'_i = \text{Splat}(\mathcal{G}, K, E_i) \tag{5}$$

By comparing the rendered image I'_i with the original object crop I_i, we quantify geometry consistency as the temporal average of SSIM (T-SSIM) across all frames:

$$\text{T-SSIM} = \frac{1}{n} \sum_i \text{SSIM}(I_i, I'_i) \tag{6}$$

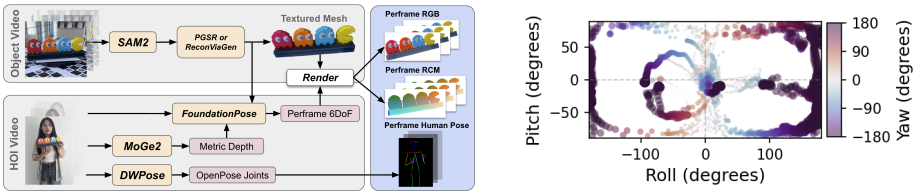
3.5 Depth-Free HOI Data Curation

The final steps of our curriculum training plan requires a carefully annotated high quality dataset. Fig. 5a shows the pipeline of our data curation pipeline for extracting all required conditions. It integrates a wide range of state-of-the-art models to achieve a robust annotation procedure:

Human Reference: if a dedicated captured human reference image is not available, we randomly choose one frame from the raw video with object masked.

Human Pose: we directly leverage DWPose [39] to infer the per frame human pose, we also filter out joints that is of too low confidence.

Object Textured Mesh: if provided with a dedicated object oriented video, first we extract object segments from the original frame sequence via SAM2,



(a) (b)

Fig. 5: (a) Data curation pipeline. (b) Rotation distribution of **Mani4D**.

Table 1: Composition of training data at different curriculum. Each curriculum consists of different combination of data annotation (P – Human Pose, H – Hand Pose, O – Textured Object Mesh).

Curriculum	Dataset	#Actors	#Objects	#Frames	P	H	O
I. Human Pose	proprietary dataset	∼1000	0	6.4M	✓	✓	
II. Hand-Object Interaction	DexYCB [4], HO3D [13], ARCTIC [9]	0, 0, 10	20, 20, 11	520K, 120K, 720K		✓	✓
Optional: Object Multiview	ScannedGoogle [7]	0	1030	550K			✓
III. Human-Object Interaction	Mani4D-Train	25*	104	45K	✓	✓	✓

similar to the practice in SeconGS [41]. We could directly infer the object mesh via an optimization based method like PGSR [6] or 2DGS [18]. We could also use diffusion based methods like ReconViaGen [3] to acquire a manifold mesh. If a dedicated object video is not available, we could also leverage Amodal3R [34] to reconstruct object from occluded segments.

Metric Depth and Scale: it could be hard and costly to acquire high quality depth map from the wild video. As an alternative, we adopt MoGe2 [31] together with camera intrinsic information extracted from image EXIF tag. Then we plainly align those predicted metric depth with RANSAC to ensure a consistent depth estimation across frames. We implant an extra scale calibration module to calibrate the scale of object mesh, inspired by OnePoseViaGen [10] and RGBTrack [12].

Object 6DoF Pose: Finally we feed the RGB image, depth map as well as the textured object mesh into FoundationPose [33] to acquire a robust object 6DoF estimation. one of the key strengths of our curated dataset is its "multi-view" nature. Fig. 5b visualizes the diverse distribution of object rotations across frames relative to the first frame, demonstrating broad coverage of object viewpoints in our training data.

Object RCM: it is relatively easy to generate the relative coordinate map of the mesh as long as the object 6DoF and mesh is given. As introduced in Sec. 3.2, we could simply recolor each vertex of the mesh with its relative coordinate, the face coloring is linearly interpolated automatically in most modern shaders.

4 Experiments

4.1 Training Details

Tab. 1 shows the composition of the training data for each curriculum mentioned in Sec. 3.3. We use Wan2.1 [28] as our base model and implemented input latent fusion for condition injection with a MLP. Specifically, we first train model on a large proprietary dataset with human pose annotation extracted by DWPose [39]. The dataset includes over 100K video clips with diverse human identities, so that our model could pick up a robust pose following capability.

Further we trained the model on a smaller yet diverse hand-object interaction dataset, components of which includes DexYCB [4], HO3D [13], and ARCTIC [9], all of which are annotated with both textured 3D mesh, object pose, as well as high quality hand pose labeling. We rendered the object RCM conditions through the method introduced in Sec. 3.2. Through this stage, our model enhances its capability of illustrating the contacts and occlusion between hands and objects.

We also synthesize an object-only dataset where a rigid object freely yet smoothly drifts within the canvas, rendered from ScannedObject [7] and backgrounded by random sampled interior room image from IKEA indoor dataset [27]. It formats the conditional video generation task as a novel view synthesis task and we hope it could encourage our model to generate cross-view consistent object renderings. However, we found the yield from this training task is marginal hence it is an optional curriculum in our final setup.

In the last stage, Curriculum III, we carefully curate a small set of videos captured by professional filming studio, named **Mani4D-Train**, where actors are showcasing a wide range of daily objects and rotate and flip those objects simultaneously. We derive the textured mesh and infer object 6DoF following the practice mentioned in Sec. 3.5. Finally, we finish up our curriculum training via this small human-object interaction dataset with comprehensive conditioning. More training detail could be find in our supplementary material.

All training is performed on A100 clusters with linear learning rate warm up till 1e-5. For RCM, we use 9 views via random sampling as the RCM-cache, considering the number of images should follow a $4k + 1$ requirement for Wan-VAE [28]. 5 suffice for simple geometry (*e.g.*, boxes) but may struggle with complex shapes; >9 yield diminishing returns.

4.2 Baseline and Metrics

Though AnchorCrafter [36] open-source testset to quantify the model's general capacity of generating Human-Object Interaction videos yet we need a testset that has more rich and abundant object manipulation like rotation and flipping to show the different view point of the object instead of plainly showing the front view, emphasizing the 3D nature of our model. Further the AnchorCrafter dataset does not provide reliable 3D mesh, neither camera intrinsics nor metric depths that are essential for our RCM condition rendering. Under this motivation, we split a small portion from our curated dataset and make

Table 2: Quantitative results on Mani4D-Test. MEt3R and T-SSIM evaluate multiview geometry consistency via feature similarity [1] and 3DGS reconstruction [44] respectively. Lower half shows the curriculum ablation study.

Method	Obj-IoU↑	Obj-CLIP↑	Face-Cos↑	LMD↓	Subj-Cons↑	Back-Cons↑	Mot-Smth↑	MEt3R↓	T-SSIM↑
Ground Truth	–	–	–	–	–	–	–	0.0419	0.9218
UniAnimate-DiT [32]	0.4644	0.7655	0.7920	0.3260	**0.9634**	**0.9495**	**0.9940**	0.0524	0.9101
MimicMotion [43]	0.4647	0.7455	0.7391	0.2017	0.9521	0.9343	0.9923	0.0638	0.9010
AnchorCrafter [36]	0.6461	0.7355	0.5771	0.2629	0.9571	0.9413	0.9926	0.0552	0.9079
Ours (I + II + III)	**0.8288**	**0.9100**	0.8891	**0.1427**	0.9535	0.9289	0.9930	0.0454	0.9147
I + III	0.7627	0.8829	0.8648	0.2054	0.9499	0.9371	0.9926	0.0569	0.9048
I + Obj + III	0.7689	0.8901	**0.8952**	0.1930	0.9498	0.9287	0.9929	0.0492	0.9105
I + II + Obj + III	0.7770	0.8946	0.8947	0.1861	0.9505	0.9287	0.9926	**0.0429**	**0.9183**

Table 3: Quantitative ablation study for RCM.

Ablation	Obj-IoU↑	Obj-CLIP↑	LMD↓	MEt3R↓	T-SSIM↑
without RCM	0.6619	0.8639	0.1878	0.0540	0.9063
with RCM	0.8288	0.9100	0.1427	0.0454	0.9147

Table 4: Quantitative result with novel human reference.

Method	Obj-IoU↑	Obj-CLIP↑	Face-Cos↑	LMD↓
AnchorCrafter [36]	0.2435	0.6894	0.3700	0.5797
Ours	0.8212	0.8957	0.7661	0.1808

Fig. 6: Qualitative comparison with prior work.

it **Mani4D-Test**. There are totally 5 unique human references and 15 test sequences, with each sequence has a unique object rotating from 90° up to 180°. 6 objects among them have never shown up in our training set and none of the human-object combination has been shown in training set as well. Each sequence is of 97 frames, approximately 6 seconds under 15 FPS.

We follow the metrics setup as previous works [36], integrating metrics including Object-IoU, Object-CLIP for object fidelity, Face-Cos-Similarity, LMD to quantify human and pose observance, and VBench [19] for overall generated video quality.

Aside of all those video generation metrics, we use T-SSIM proposed in Sec. 3.4 to quantify the object geometry consistency across the frame. Though T-SSIM and Obj-CLIP share a similar form of temporal averaging of pairwise errors. Yet Obj-CLIP measures the deviation of generated video to the ground truth video, while the T-SSIM measures the geometry inconsistency among the generated frames themselves and no ground truth frame is required.

For baseline, we select AnchorCrafter [36] as our main baseline, which is the only open-sourced method available in the research community. HunyuanVideo-HOMA [20] and DreamActor-H1 [29] are not publicly accessible, so we also compare with MimicMotion [43] and UniAnimate-DiT [32] which is a pose driven diffusion model that also based on Wan2.1 [28]. To inject the object information for other baseline methods, we use the first frame of ground truth video where the actor holds the target object. The first frame provides the full object front view and serves as the conditioning of both human and object reference. For AnchorCrafter [36], we generate all its required inputs including hand mesh and depth map. Both our method and AnchorCrafter [36] are provided with human reference image instead of GT first frame.

4.3 Quantitative Results

Tab. 2 reports the quantitative results on **Mani4D-Test**. Our method surpasses all baselines in object fidelity (Obj-IoU, Obj-CLIP), hand accuracy (LMD), and maintains competitive human similarity (Face-Cos). For overall video quality, we achieve comparable VBench scores with a slight fall back in background consistency, which we attribute to color shifting during multi-segment inference. For multiview geometry consistency, our T-SSIM closely approaches the ground truth, and we achieve the best MEt3R [1] score among all methods, confirming that our generated objects maintain consistent 3D geometry across frames.

4.4 Qualitative Results

We conducted a qualitative comparison with baselines, as shown in Fig. 6. Our approach produces consistently superior results. MimicMotion [43] fails to preserve the geometry of the reference object, as evidenced by distorted shapes highlighted in the bounding boxes. AnchorCrafter [36] suffers from severe overfitting (SpongeBob characters appearing in generated frames, despite never being conditioned on in our setup) and does not generalize to unseen input references.

UniAnimate-DiT [32] generates visually plausible videos but cannot faithfully reveal novel viewpoints of the reference object due to the absence of effective multi-view information injection.

4.5 Ablation Studies

We show the ablation of our curriculum training in the lower part of Tab. 2. Where I, II, Obj, III represents human pose curriculum, hand-object curriculum, optional object NVS curriculum and Mani4D finetuning respectively.

Introducing hand-object dataset brings a guaranteed enhancement for the hand illustration as well as object fidelity. Yet face-cos-similarity seems is of little fluctuation. Adding pure object NVS training task to the curriculum does help the object geometry consistency and full curriculum with object task achieves the best consistency score. However, such yield always comes with a deterioration of the hand interaction illustration.

Further there is a noticeable performance drop from I+II+III to I+II+Obj+III, which could be attributed for two reasons: first the gaining from pure object curriculum is marginal and could be fully supported by hand-object curriculum alone. Second, the data distribution of our synthetic NVS task is dramatically different from the authentic human object interaction setup, which has the potential to deteriorate the model performance especially the hand fidelity. So in our practice we set pure object NVS task as an optional curriculum.

We also ablate the contribution of RCM mechanism and the result is shown in Tab. 3. We train a fresh model that uses object multiview images only since Stage 2, providing a fair comparison of the RCM conditioning mechanism. The comprehensive quality enhancement fully demonstrates the effectiveness of RCM mechanism and Fig. 6 also gives a straightforward comparison.

4.6 Infer on Novel Inputs

To fully demonstrate that our model learns to generate plausible human-object interaction instead of overfitting **Mani4D** dataset, we also perform video inference in novel inputs, including novel human reference and novel object. Comprehensive comparisons and demo videos could be checked in the project page.

Novel Human. Tab. 4 shows the quantitative result when inferring the video on novel human references, all of which are generated by image generation model. Fig. 7a provides a qualitative demonstration. From result we could see that our method surpasses AnchorCrafter [36] even when generating with unseen human reference, meanwhile maintaining a decent hand accuracy (LMD).

Novel Object. We also want to make sure our model could easily generalize to other object and geometry, hence we infer the video with both novel human references and novel objects. From Fig. 7b, we could find that our model could successfully illustrate the contact and occlusion even it has never seen the human reference neither the object reference before. Yet when the mismatch between the object geometry and human pose condition is severe, there are obvious penetration effects observed.

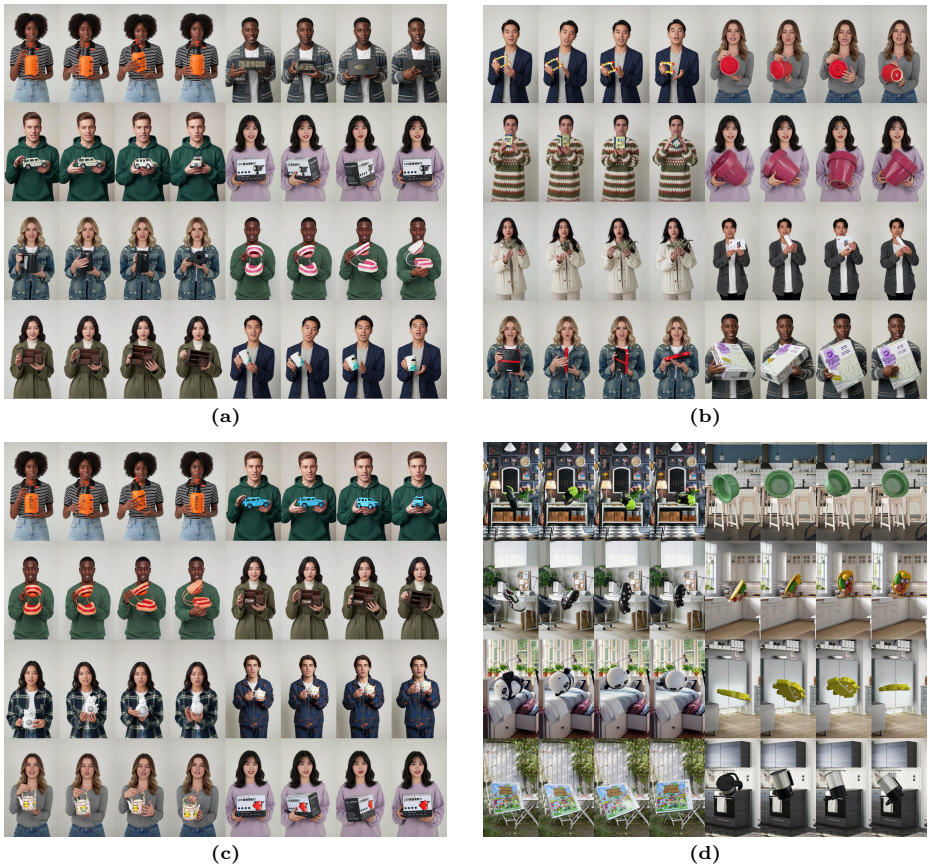


Fig. 7: Additional qualitative results. (a) Inference with novel human reference. (b) Inference with novel object and novel human reference. (c) **ByteLOOM** also provide a convenient way for object editing (d) NVS: infer with object-only condition.

Object Texture Manipulation. Since we condition the object by rendering sparse views of the 3D mesh. We could easily manipulate the object appearance via directly modifying the mesh texture. Fig. 7c shows such efforts including doodling and color toning. The generated video faithfully reflects the texture modification as we expected.

Object Novel View Synthesis. Fig. 7d tries to perform the task we previously mentioned as object-only curriculum, yet in a zero-shot style (*i.e.* with no training). The qualitative result successfully demonstrates the novel view synthesis capability facilitated by our RCM mechanism. To our surprise, the object in the generated video could also interact with the background environment in a natural and smooth way (left column in the third row in Fig. 7d).

5 Limitations and Future Work

While **ByteLOOM** advances geometry-consistent HOI video generation, several limitations remain.

First, our demonstrated setting is intentionally scoped to allocentric, studio-style manipulation rather than the full breadth of human-object interaction such as functional object use, multi-object scenes, egocentric or non-frontal viewpoints, and full-body interactions. Extending **ByteLOOM** to these in-the-wild and dexterous settings is an important direction for future work.

Second, the RCM representation assumes a rigid object with stable mesh topology; for non-rigid, articulated, or deformable objects the vertex-to-texture correspondence no longer holds, and supporting such objects (*e.g.* via part-level or articulation-aware coordinate maps) remains open.

Third, our framework conditions on geometric and texture cues rather than explicit physics, so it does not guarantee physical plausibility: when the object geometry and human pose conditions are poorly matched, penetration or contact-free artifacts can still arise, and grasp affordance is learned only implicitly from data. Incorporating contact and physical priors to enforce plausible grasping is a promising avenue.

We envision **ByteLOOM** as a solid step toward controllable, geometry-aware HOI generation, with these directions outlining a clear path forward.

6 Conclusion

We present **ByteLOOM**, a DiT-based framework for generating geometrically consistent HOI videos. Our approach addresses two critical challenges: limited viewpoint diversity of manipulated objects and dependency on fine-grained hand mesh annotations. Through the RCM-cache mechanism, we achieve precise 6-DoF object control using relative coordinate maps as a universal representation, substantially enhancing cross-frame geometry consistency. Our progressive three-stage training curriculum effectively overcomes data scarcity while eliminating hand mesh requirements. Combined with our depth-free data curation pipeline, this enables scalable and robust extraction of HOI conditioning signals. We also introduce **Mani4D**, a new benchmark for evaluating geometry-aware HOI generation capabilities. Extensive experiments demonstrate that **ByteLOOM** achieves competitive performance in human identity preservation and motion smoothness while significantly surpassing prior methods in object manipulation quality and multi-view consistency.

Acknowledgements

This work was conducted as part of a summer internship project at ByteDance. Authors, Bangya Liu and Suman Banerjee were also partially supported the U.S. National Science Foundation under grants 2504963, 2312716, 2212688, 2112562, and 2107060.

References

1. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6034–6044 (2025) 8, 11, 12
2. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019) 5
3. Chang, J., Ye, C., Wu, Y., Chen, Y., Zhang, Y., Luo, Z., Li, C., Zhi, Y., Han, X.: Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. arXiv preprint arXiv:2510.23306 (2025) 9
4. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9044–9053 (2021) 9, 10
5. Chen, B., Zhong, C., Xiang, W., Geng, Y., Xie, X.: Virtualmodel: Generating object-id-retentive human-object interaction image by diffusion model for e-commerce marketing. arXiv preprint arXiv:2405.09985 (2024) 3
6. Chen, D., Li, H., Ye, W., Wang, Y., Xie, W., Zhai, S., Wang, N., Liu, H., Bao, H., Zhang, G.: Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction (2024) 9
7. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022) 9, 10
8. Fan, Y., Yang, Q., Wang, K., Zhou, H., Li, Y., Feng, H., Ding, E., Wu, Y., Wang, J.: Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. CVPR (2025) 4
9. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12943–12954 (2023) 9, 10
10. 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. arXiv preprint arXiv:2509.07978 (2025) 9
11. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–12 (2025) 4
12. Guo, T., Yu, J.: Rgbtrack: Fast, robust depth-free 6d pose estimation and tracking. arXiv preprint arXiv:2506.17119 (2025) 9
13. Hampali, S., Sarkar, S.D., Lepetit, V.: HO-3D_v3: Improving the accuracy of hand-object annotations of the HO-3D dataset. arXiv preprint arXiv:2107.00887 (2021) 9, 10
14. He, H., Liang, Y., Wang, L., Cai, Y., Xu, X., Guo, H.X., Wen, X., Chen, Y.C.: Lucidfusion: Generating 3d gaussians with arbitrary unposed images (2024) 4, 6
15. Hoe, J.T., Jiang, X., Chan, C.S., Tan, Y.P., Hu, W.: Interactdiffusion: Interaction control in text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6180–6189 (2024) 3

16. Hu, H., Wang, W., Zhou, W., Li, H.: Hand-object interaction image generation. Advances in Neural Information Processing Systems **35**, 23805–23817 (2022) 3
17. Hu, L., Wang, G., Shen, Z., Gao, X., Meng, D., Zhuo, L., Zhang, P., Zhang, B., Bo, L.: Animate anyone 2: High-fidelity character image animation with environment affordance. arXiv preprint arXiv:2502.06145 (2025) 4
18. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024) 9
19. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024) 12
20. Huang, Z., Zhou, Z., Cao, J., Ma, Y., Chen, Y., Rao, Z., Xu, Z., Wang, H., Lin, Q., Zhou, Y., et al.: Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. arXiv preprint arXiv:2506.08797 (2025) 2, 4, 5, 7, 12
21. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025) 5
22. Men, Y., Yao, Y., Cui, M., Bo, L.: Mimo: Controllable character video synthesis with spatial decomposed modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21181–21191 (2025) 4
23. Pang, Y., Shao, R., Zhang, J., Tu, H., Liu, Y., Zhou, B., Zhang, H., Liu, Y.: Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12209–12219 (2025) 4
24. Shi, Y., Wang, P., Ye, J., Long, M., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023) 4
25. Song, W., Zhang, X., Li, S., Gao, Y., Hao, A., Hou, X., Chen, C., Li, N., Qin, H.: Hoianimator: Generating text-prompt human-object animations using novel perceptive diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 811–820 (2024) 3
26. Song, Z., Gong, X., Liu, B., Zhao, Z.: Mv-s2v: Multi-view subject-consistent video generation (2026), https://arxiv.org/abs/2601.17756 4
27. Tautkute, I., Możejko, A., Stokowiec, W., Trzciński, T., Brocki, Ł., Marasek, K.: What looks good with my sofa: Multimodal search engine for interior design. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Proceedings of the 2017 Federated Conference on Computer Science and Information Systems. Annals of Computer Science and Information Systems, vol. 11, pp. 1275–1282. IEEE (2017) 10
28. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 10, 12
29. Wang, L., Xia, Z., Hu, T., Wang, P., Wei, P., Zheng, Z., Zhou, M., Zhang, Y., Gao, M.: Dreamactor-h1: High-fidelity human-product demonstration video generation

via motion-designed diffusion transformers. arXiv preprint arXiv:2506.10568 (2025) 2, 4, 5, 7, 12
30. Wang, Q., Luo, Y., Shi, X., Jia, X., Lu, H., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–10 (2025) 4
31. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025) 9
32. Wang, X., Zhang, S., Gao, C., Wang, J., Zhou, X., Zhang, Y., Yan, L., Sang, N.: Unianimate: Taming unified video diffusion models for consistent human image animation. Science China Information Sciences **68**(10), 1–14 (2025) 1, 11, 12, 13
33. Wen, B., Yang, W., Kautz, J., Birchfield, S.: FoundationPose: Unified 6d pose estimation and tracking of novel objects. In: CVPR (2024) 9
34. Wu, T., Zheng, C., Guan, F., Vedaldi, A., Cham, T.J.: Amodal3r: Amodal 3d reconstruction from occluded 2d images. arXiv preprint arXiv:2503.13439 (2025) 9
35. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14928–14940 (2023) 3
36. Xu, Z., Huang, Z., Cao, J., Zhang, Y., Cun, X., Shuai, Q., Wang, Y., Bao, L., Li, J., Tang, F.: Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. arXiv preprint arXiv:2411.17383 (2024) 2, 4, 5, 7, 10, 11, 12, 13
37. Xue, M., Liu, Y., Guo, L., Huang, S., Ding, C.: Guiding human-object interactions with rich geometry and relations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22714–22723 (2025) 3
38. Xue, Z.S., Luo, R., Chen, C., Grauman, K.: Hoi-swap: Swapping objects in videos with hand-object interaction awareness. Advances in Neural Information Processing Systems **37**, 77132–77164 (2024) 4
39. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 8, 10
40. Yi, H., Thies, J., Black, M.J., Peng, X.B., Rempe, D.: Generating human interaction motions in scenes with text control. In: European Conference on Computer Vision. pp. 246–263. Springer (2024) 3
41. Yin, H., Zhan, H., Xu, Y., Yeh, R.A.: Semantic consistent language gaussian splatting for point-level open-vocabulary querying. arXiv preprint arXiv:2503.21767 (2025) 9
42. Zhang, H., Chen, X., Wang, Y., Liu, X., Wang, Y., Qiao, Y.: 4diffusion: Multi-view video diffusion model for 4d generation. Advances in Neural Information Processing Systems **37**, 15272–15295 (2024) 4
43. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv preprint arXiv:2406.19680 (2024) 1, 11, 12
44. Zhao, Q., Tan, H., Wang, Q., Bi, S., Zhang, K., Sunkavalli, K., Tulsiani, S., Jiang, H.: E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. arXiv preprint arXiv:2512.10950 (2025) 8, 11

45. Zhao, Z., Fan, F., Liao, W., Yan, J.: Grounding and enhancing grid-based models for neural fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19425–19435 (June 2024) 4
46. Zhao, Z., Gong, X., Liu, B., Song, Z., Zhang, J., Wu, S., Chen, Y., Zhang, H.: Cetcam: Camera-controllable video generation via consistent and extensible tokenization. arXiv preprint arXiv:2512.19020 (2025) 4, 5
47. Zhao, Z., Molodyk, P., Xue, H., Chen, Y.: Laplacian multi-scale flow matching for generative modeling (2026), https://arxiv.org/abs/2602.19461 3
48. Zhou, J., Wu, Y., Li, S., Wei, M., Fan, C., Chen, W., Jiang, W., Wang, F.: Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. arXiv preprint arXiv:2504.14977 (2025) 1, 5