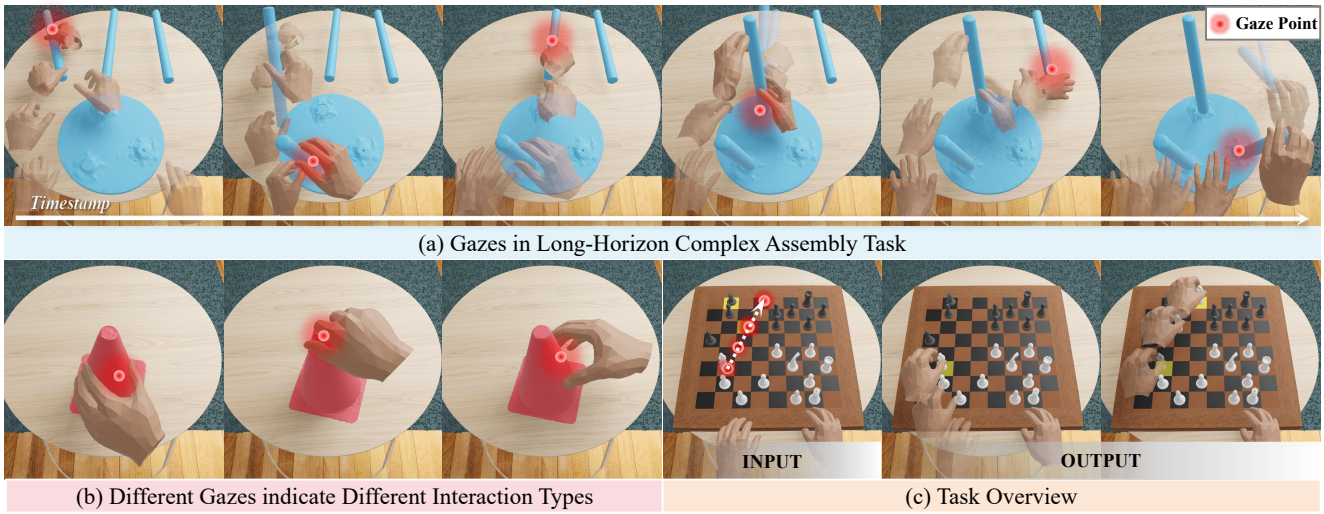
Gaze-guided Hand-Object Interaction Synthesis: Dataset and Method

Jie Tian, Ran Ji, Lingxiao Yang, Suting Ni, Yuexin Ma *Member, IEEE*, Lan Xu *Member, IEEE*, Jingyi Yu *Fellow, IEEE*, Ye Shi *Member, IEEE*, Jingya Wang *Member, IEEE*



(a) Gazes in Long-Horizon Complex Assembly Task

(b) Different Gazes indicate Different Interaction Types

(c) Task Overview

Figure 1. Gaze plays a crucial role in hand-object interaction tasks. The subfigure (a) illustrates how gaze helps coordinate the hands, objects, and brain during complex table assembly. The subfigure (b) shows how gaze affects attention allocation and grasp strategies. Building on these observations, we introduce the gaze-guided hand-object interaction synthesis task, as depicted in the subfigure (c), which uses gaze data as input to generate hand-object interactions that align with human intentions.

arXiv:2403.16169v6 [cs.CV] 17 Mar 2026

***Abstract*—Gaze plays a crucial role in revealing human attention and intention, particularly in hand-object interaction scenarios, where it guides and synchronizes complex tasks that require precise coordination between the brain, hand, and object. Motivated by this, we introduce a novel task: Gaze-Guided Hand-Object Interaction Synthesis, with potential applications in augmented reality, virtual reality, and assistive technologies. To support this task, we present GazeHOI, the first dataset to capture simultaneous 3D modeling of gaze, hand, and object interactions. This task poses significant challenges due to the inherent sparsity and noise in gaze data, as well as the need for high consistency and physical plausibility in generating hand and object motions. To tackle these issues, we propose a stacked gaze-guided hand-object interaction diffusion model, named GHO-Diffusion. The stacked design effectively reduces the complexity of motion generation. We also introduce HOI-Manifold Guidance during the sampling stage of GHO-Diffusion, enabling fine-grained control over generated motions while maintaining the data manifold. Additionally, we propose a spatial-temporal gaze feature encoding for the diffusion condition and select diffusion results based on consistency scores between gaze-contact maps and gaze-interaction trajectories. Extensive experiments highlight the effectiveness of our method and the unique contributions of our dataset. More details in https://takiee.github.io/gaze-hoi/.**

Received 07 Jan 2025; revised 14 Dec 2025; accepted 28 Jan 2026. This work was supported by National Natural Science Foundation of China (62406195, 62303319), ShanghaiTech AI4S Initiative SHTAI4S202404, HPC Platform of ShanghaiTech University, and MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), and Shanghai Engineering Research Center of Intelligent Vision and Imaging. This work was also supported in part by computational resources provided by Fcloud CO., LTD.

The authors are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: {tianjie2022, jiran2022, yanglx2023, nist2024, mayuexin, xulan1, yujingyi, shiye, wangjingya}@shanghaitech.edu.cn).Corresponding author: Jingya Wang.

***Index Terms*—Hand-object interaction, gaze, dataset, motion generation.**

I. Introduction

GAZE serves as a significant behavioral signal, directly reflecting human attention distribution and cognitive processes [1]–[3]. Extensive research has been conducted on gaze, including gaze estimation [4]–[10] and human action prediction based on gaze [11]–[15]. While these studies have explored gaze in various applications, they have primarily focused on broad human motions, with limited attention given to hand movements and detailed dynamic interactions within fine-grained environments. Furthermore, gaze is particularly crucial in hand-object interactions, as it reveals "*when, where, and how*" the user interacts with the object. Figure 1(a) illustrates how gaze supports long-horizon synthesis for complex tasks, while Figure 1(b) shows that gaze indicates how humans interact with objects using different parts and interaction types. By effectively identifying objects and areas of focus, gaze enables the seamless integration of human intentions and motions, ensuring precise coordination and consistency between the brain, hands, and objects. This coordination is essential for executing complex interactions, such as grasping, manipulating, or positioning objects, thereby enhancing both the efficiency and accuracy of these motions.

Motivated by this, we propose a novel task: Gaze-guided Hand-Object Interaction Synthesis, as shown in Figure 1(c). This task leverages gaze to naturally generate motions that align with human intent, eliminating the need for additional

textual annotations or motion references. This approach is particularly well-suited for integration into VR/AR environments and assistive technologies. For example, in VR/AR settings, gaze-guided technology allows users to interact with virtual objects more intuitively by accurately predicting their intentions without the need for physical controllers. Additionally, it benefits individuals with disabilities by allowing them to control interactions through eye movements, enhancing their ability to express intentions and perform tasks

Existing hand-object datasets [16]–[25] are limited to interactions between hands and objects, often neglecting gaze information. To fill this gap, we introduce GazeHOI, the first dataset that simultaneously integrates 3D modeling of gaze, hand, and object interactions. GazeHOI presents significant challenges due to the diverse shapes and sizes of objects, ranging from small items like chess pieces to larger, assemblable furniture. Additionally, the dataset covers a wide range of complex tasks, including repositioning objects, selecting specific targets within cluttered environments, organizing disordered items, and assembling furniture.

However, leveraging gaze data for generating hand and object motions presents significant challenges. The inherent sparsity and noise in gaze data make it difficult to reliably capture and interpret user intentions. Additionally, achieving high consistency and physical plausibility in the generated motions is a complex task, as it requires accurate synchronization between gaze input and corresponding hand-object interactions. To address these challenges, we propose a novel approach: a stacked gaze-guided hand-object interaction diffusion model, named GHO-Diffusion. We decouple the task into two stages: gaze-activated object dynamics synthesis and object-driven hand kinematic synthesis. Moreover, we introduce HOI-Manifold Guidance during the sampling stage of GHO-Diffusion. This technique allows for fine-grained control over the generated motions, ensuring they remain within the data manifold and adhere to the natural constraints of hand-object interactions. Additionally, to further leverage gaze information, we propose a spatial-temporal gaze feature encoding approach for the diffusion condition. This method captures both spatial and temporal aspects of gaze data, enhancing the diffusion process with detailed representations of gaze interactions. We also select diffusion results based on consistency scores between gaze-contact maps and gaze-interaction trajectories. Our contributions can be summarized as follows:

1) A Novel Task and Dataset. We introduce the novel task of gaze-guided hand-object interaction synthesis, which utilizes gaze to generate natural interaction motions that align with human intent. To support this task, we present GazeHOI, the first dataset incorporating gaze information for hand-object interactions.

2) GHO-Diffusion Model. We develop the Gaze-guided Hand-Object interaction Diffusion model, GHO-Diffusion, which employs stacked diffusion models to reduce the complexity of motion modeling. Furthermore, we integrate the HOI-Manifold Guidance during the inference phase of GHO-Diffusion, providing fine-grained control over generated motions while maintaining the data distribution learned by the diffusion model.

3) Gaze-Guided Feature Encoding and Motion Selection. We propose a gaze-guided spatial-temporal feature encoding as the diffusion condition, coupled with a mechanism to select diffusion results based on consistency scores between gaze-contact maps and gaze-interaction trajectories. This approach ensures that the generated motions are precisely aligned with the given gaze input.

TABLE I
COMPARISON WITH EXISTING GAZE DATASETS

	Gaze	Object pose	Hand pose
MoGaze [11]	✓	✓	×
GIMO [12]	✓	×	✓
ActionSense [13]	✓	×	✓
MECCANO [14]	✓	×	×
HoloAssist [28]	✓	×	✓
ADT [29]	✓	✓	×
EgoBody [30]	✓	×	✓
GazeHOI	✓	✓	✓

II. Related Work

A. Gaze for Perception

Gaze is a key behavioral signal for inferring intentionality, attention, and cognitive processes. Early research, limited by equipment accuracy, focused on third-person gaze estimation [4]–[10] to predict human attention regions. Recent tracking advancements [27] have facilitated diverse first-person datasets [11]–[15], [28]–[30] that leverage gaze for contextual tasks like ego-trajectory inference [15] or action labeling [31]. However, existing gaze-motion benchmarks [11]–[13], [28]–[30] largely overlook the link between gaze and hand-object interaction (HOI)—essential for fine-grained intent inference and realism in VR. Specifically, current datasets omit either hand [11], [29] or object poses [12], [13], [28], [30], as shown in Table I. Methodologically, most studies focus on human motion prediction [12], [32], [33], treating gaze as auxiliary input to fusion encoders and relying heavily on historical sequences. These approaches fail to model HOI precision, such as realistic contact, penetration avoidance, and plausible physical dynamics. To bridge these gaps, we present the first dataset capturing 3D gaze, hand, and object interactions simultaneously. We introduce a novel task: gaze-guided HOI synthesis from an initial state without relying on historical motions. Our method utilizes gaze-guided spatial-temporal feature encoding as a diffusion condition and selects results based on gaze-contact and trajectory consistency scores to effectively capture the intentionality inherent in gaze.

B. Hand-Object Interaction Dataset

Numerous datasets [16]–[25] dedicated to hand-object interactions have been developed to facilitate the capture of hand and object poses and to deepen our understanding of these interactions, as shown in Table II. Most datasets only capture simple interaction patterns,like grasping and moving, which is common but not comprehensive enough. HOI4D [20] and ARCTIC [21] explore the interaction of articulated objects. TACO [22] and OakInk2 [25] captures scenes of

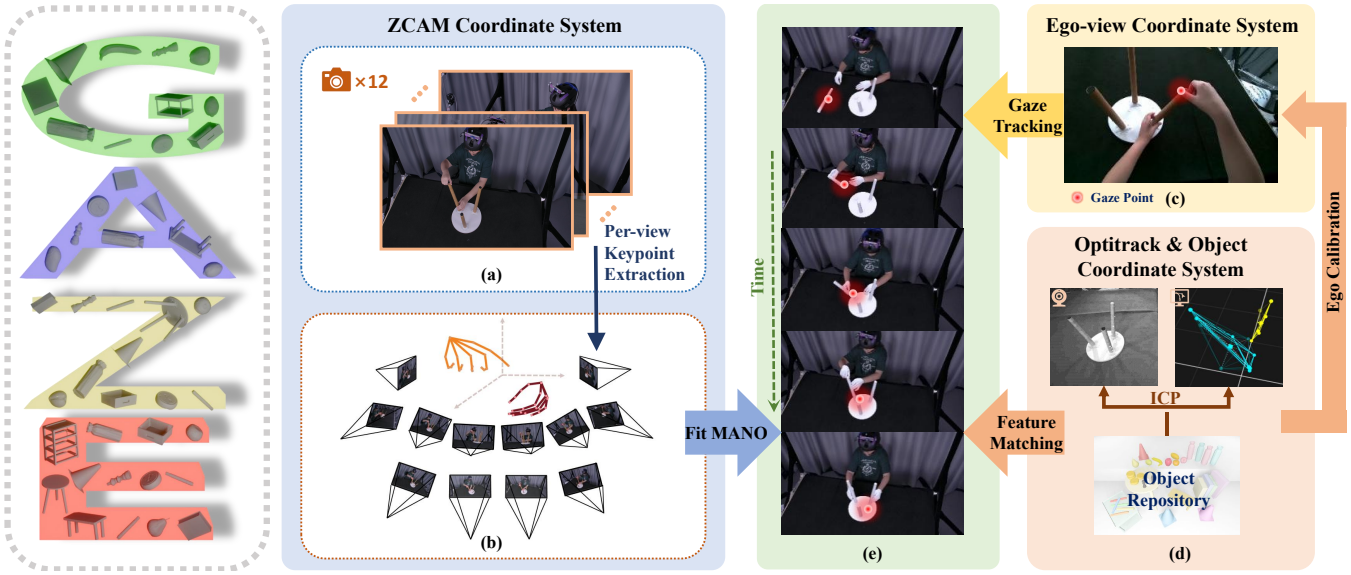


Fig. 2. Automatic data processing pipeline: (a) raw 12-view images; (b) 3D hand joint estimation via MediaPipe [26] and triangulation; (c) ego-view gaze; (d) object scanning and marker tracking; (e) final hand-object motion.

TABLE II
COMPARISON WITH EXISTING 4D HAND-OBJECT INTERACTION DATASETS.

	Gaze	Ego	Multi-obj	Complex	#views	#images
HO-3D [16]	×	×	×	×	1-4	78K
DexYCB [17]	×	×	×	×	8	582K
H2O [18]	×	✓	×	×	5	571K
OakInk [19]	×	×	×	×	4	230K
HOI4D [20]	×	✓	×	×	2	2.4M
ARCTIC [21]	×	✓	×	×	9	2.1M
TACO [22]	×	✓	✓	✓	13	5.2M
OAKINK2 [25]	×	✓	✓	✓	4	4.01M
GazeHOI	✓	✓	✓	✓	13	3.2M

simultaneous interactions with multiple objects. Furthermore, existing datasets focus exclusively on the relationship between hands and objects, neglecting the role of gaze. To address this gap, we introduce a comprehensive egocentric dataset enriched with detailed gaze annotations. This dataset captures the dynamics of hand-object interactions across a wide range of tasks, from arranging chess pieces to assembling furniture — scenarios not covered in existing datasets. Our goal is to bridge the shortcomings of current datasets by providing a resource that reflects the complexity of real-world interactions and emphasizes the importance of gaze in predicting human intent and managing multi-object manipulation.

C. Hand-Object Motion Synthesis

In the field of hand-object motion synthesis, prior efforts have mainly centered on frame-wise generation [34]–[37]. For dynamic hand-object interaction, previous methods typically rely on motion references, such as trajectories [38]–[43] pr grasp references [44], or only focus on approaching stage [45], [46]. Recently, there are several works that study the hand-object interaction generation directly based on text [47], [48], significantly reducing the reliance on specific conditions. Building on advancements in 3D gaze acquisition technology, our work further relaxes these constraints by generating hand-object interactions directly from gaze data. Compared to text, gaze-based generation aligns more closely with the user's intent, resulting in more intuitive hand-object interactions.

III. GAZEHOI DATASET

Existing datasets primarily focus on the interaction between hands and objects, overlooking the crucial role of gaze. However, a comprehensive understanding of interaction requires coordinated hand-eye-object cooperation. To address this gap, we present GazeHOI, the first interaction dataset that simultaneously incorporates 3D modeling of gaze, hand, and object interactions. GazeHOI encompasses 237K frames with an average duration of 19.1 seconds and 2-4 objects per sequence, further divisible into 1378 subsequences. We collect 33 different objects with various shape and size, the smallest of which is a chess piece with a height of 7cm and a mere diameter of 2cm. Each sequence is accompanied by a task-level description, which varies in complexity, and includes repositioning objects to another location, selecting specific targets from cluttered environments, organizing items from disarray into order, and assembling furniture. This rich variety aims to provide comprehensive insights into the intricate dynamics of interactive behaviors.

A. Dataset Hardware Setup

To construct the GazeHOI dataset, we developed a data collection platform encompassing a 1.6m × 1.2m area, equipped with a multi-camera system. This platform comprises four distinct equipment systems. The ZCAM system includes 12 ZCAM E2 cameras that have been synchronized through hardware, capturing RGB signals at a resolution of 3840 x 2160 at 30fps, which are utilized for subsequent hand pose acquisition. The Optitrack system consists of 8 hardware-synchronized Optitrack px13W cameras, which use infrared to capture the positions of reflective markers at a frame rate of 120fps, facilitating the tracking of object poses. The Scanner system, comprising two EinScan Pro 2X and EinScan SE scanners, is dedicated to collecting geometric information about

objects. The Ego system, primarily aimed at gathering gaze data, employs a Pupil Core connected to a RealSense D455, simultaneously recording gaze information along with RGBD data from an ego-centric perspective. We devise a sophisticated workflow to achieve coordinate system registration among the 4 equipment systems. Each system is tasked with collecting specific types of data, and we have devised a sophisticated workflow to achieve coordinate system registration among the 4 equipment systems through reflective markers. We attached eight 10mm diameter markers on the surface of the RealSense and scanned it to obtain its 3D model. By annotating the position of each marker in every RGB viewpoint and triangulating these positions, we could ascertain the location of markers within each system, thereby aligning the coordinate systems of multiple devices.

B. Data Collection

We recruited 10 volunteers (5 males, 5 females) and provided them with comprehensive scripts detailing each task. Each task was performed over multiple iterations to capture natural gaze transitions. In unfamiliar settings, participants tend to scan their surroundings to understand the environment, leading to longer gaze durations on task-relevant objects. As familiarity increases, gaze becomes more focused and durations decrease. This protocol ensures the authenticity of our data by accurately reflecting real-world gaze behavior and adaptation during goal-oriented interactions.

C. Data Annotation

1) 3D Hand Pose: 12 synchronized ZCAM cameras provide an exhaustive 3D perspective of the hand (Fig. 2a). We track 2D hand keypoints via MediaPipe [26], which are then triangulated into 3D global coordinates (Fig. 2b). Finally, we optimize MANO [49] parameters by enforcing joint alignment and realistic anatomical constraints.

2) Object 6D Pose: We focus on rigid objects, determining 6D poses by tracking at least eight 3mm reflective markers per object via Optitrack. Markers are strategically placed to minimize interaction interference and occlusion. After capturing precise geometries via a 3D scanner, we employ the Iterative Closest Point (ICP) algorithm to convert Optitrack reflection points into 6D poses (Fig. 2d).

3) Gaze Acquisition: The eye tracker Pupil Core provides 3D egocentric gaze data. To unify gaze, hand, and object data in a single coordinate system, we perform ego-calibration by attaching markers to the egocentric camera. Post-processing, including filtering short-duration gaze and smoothing, is applied to enhance data quality and reduce noise (Fig. 2c).

IV. Method

A. Problem Definition

We define a gaze-consistent hand-object interaction sequence of length l as $\mathbf{S} = \langle \mathbf{G}, \mathbf{H}, \mathbf{O} \rangle$, where $\mathbf{G} = \{\mathbf{g}_i\}_{i=1}^{l}$ represents the gaze sequence, with each $\mathbf{g}_i \in \mathbb{R}^3$ denoting the gaze point in the i-th frame. $\mathbf{H} = \{\mathbf{h}_l^i, \mathbf{h}_r^i\}_{i=1}^{l}$ denotes the hand motion sequences, comprising left-hand motions $\mathbf{h}_l^i$ and right-hand motions $\mathbf{h}_r^i$. Each hand motion $\mathbf{h}^i = \{\mathbf{R}_h^i, \mathbf{T}_h^i, \boldsymbol{\theta}^i, \boldsymbol{\beta}\} \in \mathbb{R}^{61}$ follows the MANO [49] parameterization, where $\mathbf{R}_h^i$ is the rotation, $\mathbf{T}_h^i$ is the translation, $\boldsymbol{\theta}^i$ represents the pose and $\boldsymbol{\beta}$ represents shape parameters; $\mathbf{O} = \{\mathbf{o}^i\}_{i=1}^{l}$ is the object motion sequence, where each $\mathbf{o}^i = \{\mathbf{R}_o^i, \mathbf{T}_o^i\} \in \mathbb{SE}(3)$ denotes the object's pose, with $\mathbf{R}_o^i$ as the rotation and $\mathbf{T}_o^i$ as the translation. Additionally, we define the object geometry as a set of vertices $\mathbf{P}_o = \{\mathbf{p}_o^n\}_{n=1}^{N} \in \mathbb{R}^{N\times 3}$, where N represents the number of object vertices. Given the input $\mathcal{I} = \{\mathbf{G}, \mathbf{h}_l^1, \mathbf{h}_r^1, \mathbf{o}^1, \mathbf{P}_o\}$, our objective is to derive a subsequent hand-object interaction sequence $\langle \mathbf{H}, \mathbf{O} \rangle$ that is not only consistent with the gaze but also exhibits natural and realistic interactions between the hand and the object.

B. GHO-Diffusion: Gaze-guided Hand-Object Diffusion

GHO-Diffusion decouples the complex hand-object interaction into two stacked diffusion stages, Object Dynamics Synthesis with Spatial-Temporal Gaze Encoding and Hand Kinematic Synthesis with HOI-Manifold Guidance as shown in Figure 3(a). This decoupling strategy reduces modeling complexity while effectively capturing both global spatial-temporal positioning and the intricate details of hand-object interaction motions.

The diffusion model contains a fixed forward process $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$ which gradually adds Gaussian noise to clean data $\mathbf{x}_0$ until it becomes pure noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and train a denoiser $p_\theta(\mathbf{x}_t, t, \mathbf{c})$ to gradually denoise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to generate $\mathbf{x}_0$ in the reverse process. Instead of predicting $\boldsymbol{\epsilon}_t$ at each time step t, we follow MDM [50] to predict $\hat{\mathbf{x}}_0$ and optimizing $p_\theta(\mathbf{x}_t, t, \mathbf{c})$ using the simple objective function:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\mathbf{x}_0\sim q(\mathbf{x}_0|\mathbf{c}), t\sim[1,T]} ||\mathbf{x}_0 - p_\theta(\mathbf{x}_t, t, \mathbf{c})||_2^2. \quad (1)$$

1) Stage-1: Object Dynamics Synthesis with Spatial-Temporal Gaze Encoding.: The gaze point locations are closely related to the object surface vertices, and since object motions have fewer degrees of freedom compared to hand motions, so we generate the object motions first. Gaze is a sparse and noisy signal, making it challenging to encode effectively. Drawing inspiration from GIMO [12], we design a gaze-guided spatial-temporal feature encoding module for object diffusion. We define $\mathbf{f}_o^k$ as the k-th local per-point object spatial feature extracted by PointNet++ [51]. Then according to the spatial relationship between gaze and object points, we can calculate gaze spatial feature:

$$\mathbf{f}_g = \|\mathbf{G} - \mathbf{P}_o^{k^*}\|^{-1} \mathbf{f}_o^{k^*}, \quad (2)$$

where $k^* = \arg\min_{k=1}^{N} \|\mathbf{G} - \mathbf{P}_o^k\|$, representing the index of the object points that is closest to the gaze point in i-th frame. Subsequently, these spatial features are processed through a self-attention mechanism to extract the final spatial-temporal condition features:

$$\mathbf{c}_g = \text{softmax}\left(\frac{\mathbf{f}_g\mathbf{W}_Q(\mathbf{f}_g\mathbf{W}_K)^\top}{\sqrt{d}}\right)\mathbf{f}_g\mathbf{W}_V, \quad (3)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ are trainable weight matrices, and d is the dimension of the feature.

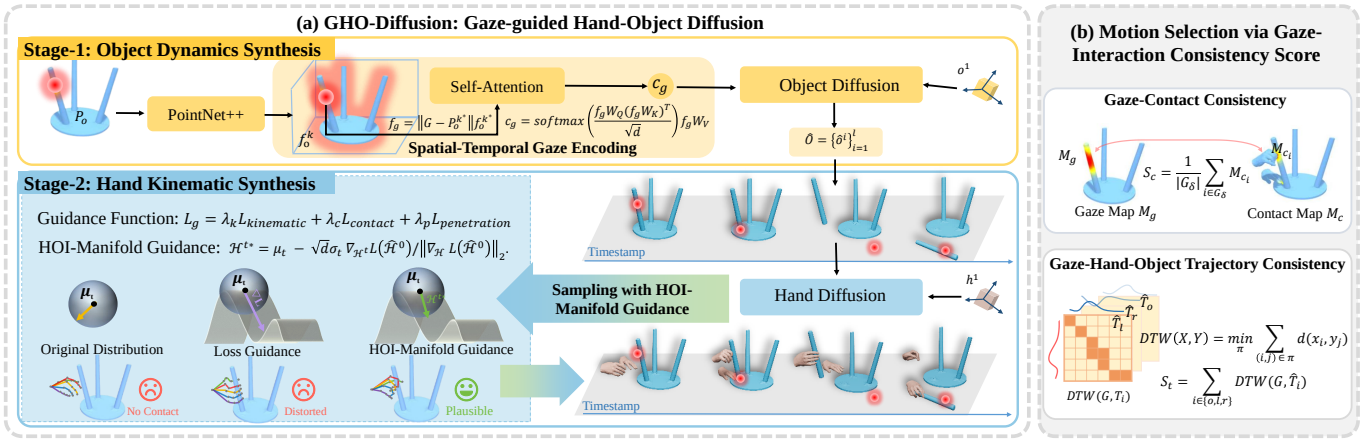


Fig. 3. Pipeline overview. The subfigure (a) illustrates the stacked framework of the gaze-guided hand-object interaction diffusion model, GHO-Diffusion. The subfigure (b) shows the gaze-interaction consistency score for motion selection.

We leverage the gaze spatial-temporal features $\mathbf{c}_g$, along with the object geometry $\mathbf{P}$ and initial object pose $\mathbf{o}^1$ as stage-1 condition and employ the objective function in Equation 4 as Stage-1 training loss to generate object motions $\hat{\mathbf{O}}$:

$$\mathcal{L}_{\text{stage1}} = \lambda_\alpha \mathcal{L}_{\text{simple}} + \lambda_v \mathcal{L}_{\text{verts}} + \lambda_t \mathcal{L}_{\text{trans}} + \lambda_s \mathcal{L}_{\text{smooth}}. \quad (4)$$

where $\mathcal{L}_{\text{simple}}$ is defined by Equation 1.

The object translation loss $\mathcal{L}_{\text{trans}}$ is defined as:

$$\mathcal{L}_{\text{trans}} = \sum_{i=1}^{L} ||\mathbf{T}_o^i - \hat{\mathbf{T}}_o^i||_2, \quad (5)$$

which enforces a constraint on the object's position. The object vertices loss $\mathcal{L}_{\text{verts}}$ is defined as:

$$\mathcal{L}_{\text{verts}} = \sum_{i=1}^{l} ||(\mathbf{R}_o^i(\mathbf{P}) + \mathbf{T}_o^i) - (\hat{\mathbf{R}}_o^i(\mathbf{P}) + \hat{\mathbf{T}}_o^i)||_2, \quad (6)$$

which enforces constraints on the object's surface geometry, promoting finer details and ensuring more precise alignment. The smoothness loss $\mathcal{L}_{\text{smooth}}$ is defined as:

$$\mathcal{L}_{\text{smooth}} = ||\hat{\mathbf{O}}^{1:l} - \hat{\mathbf{O}}^{0:l-1}||_2, \quad (7)$$

which promotes the continuity and fluidity of the generated motion. By minimizing the difference between consecutive motion frames, this loss ensures that the generated object motion is smooth and avoids abrupt transitions between frames.

2) Stage-2: Hand Kinematic Synthesis with HOI Guidance: In this stage, we leverage the object motions $\hat{\mathbf{O}}$ generated in Stage-1 to synthesize the corresponding hand motions $\hat{\mathbf{H}}$ through a hand diffusion process. Hand motion synthesis presents greater complexity compared to object motion synthesis, primarily because hand movements are highly articulated and require more stringent constraints to ensure natural and physically plausible outcomes. To address this, we adopt a canonicalized hand-object interaction representation: $\mathcal{H} = [\mathbf{J}, \mathbf{C}, \mathbf{F}, \mathbf{v}, \mathbf{a}]$. Here, hand poses are represented by joint positions $\mathbf{J}$, as neural networks typically learn translations more effectively than rotations. The contact flag $\mathbf{C} \in \mathbb{R}^{42}$ encodes the contact status of each joint, while the offset vector $\mathbf{F}$ represents both hand-object and inter-hand spatial displacements. The velocity vector $\mathbf{v}$ captures both linear and angular velocities for each hand, including their relative velocities, and the acceleration vector $\mathbf{a}$ is similarly defined.

To generate the hand motions $\hat{\mathbf{H}}$, we condition the synthesis on the generated object motions $\hat{\mathbf{O}}$, alongside object geometry $\mathbf{P}$ and the initial hand pose $\mathbf{h}^1$ in Stage-2. For training, we employ the objective function from Equation 1 and introduce a bone length loss $\mathcal{L}_{\text{bone}}$ to maintain the structural integrity of the hand:

$$\mathcal{L}_{\text{stage2}} = \lambda_\beta \mathcal{L}_{\text{simple}} + \lambda_b \mathcal{L}_{\text{bone}}. \quad (8)$$

We define $\mathcal{B}(\cdot)$as a function that calculates the bone lengths, taking the hand joints $\mathbf{J}$ as input. The bone length loss $\mathcal{L}_{\text{bone}}$ is defined as:

$$\mathcal{L}_{\text{bone}} = ||\mathcal{B}(\mathbf{J}) - \mathcal{B}(\hat{\mathbf{J}})||_2. \quad (9)$$

Since hand diffusion effectively captures the overall distribution of hand-object interactions, it falls short of achieving precision in finer details, such as maintaining contact or avoiding penetrations. Therefore, we introduce HOI-Manifold Guidance to strengthen the physics constraint while preserving the original data distribution. We define three guidance functions from the perspective of Kinematic, Contact and Penetration.

*a) **Hand-Object Kinematic Guidance.**:* We define redundancy parameters to enhance the reliability of hand motions while using only the hand joints for subsequent steps. The kinematic consistency guidance is designed for test-time adaptation, ensuring that the generated parameters remain consistent with those recalculated from the generated hand joints and object motion conditions. The loss function is as follow,

$$\mathcal{L}_{\text{kinematic}} = \|\hat{\mathbf{F}} - \tilde{\mathbf{F}}\|_2 + \|\hat{\mathbf{v}} - \tilde{\mathbf{v}}\|_2 + \|\hat{\mathbf{a}} - \tilde{\mathbf{a}}\|_2. \quad (10)$$

*b) **Hand-Object Contact Guidance.**:* We define a contact loss to minimize the distance between hand joints that are near the object's surface but not yet in contact:

$$\mathcal{L}_{\text{contact}} = \frac{\sum_{i=1}^{l} d(\hat{\mathbf{J}}^i, \mathbf{P}_o^i) \cdot \mathbb{I}(d(\hat{\mathbf{J}}^i, \mathbf{P}_o^i) < \tau))}{\sum_{i=1}^{l} \mathbb{I}(d(\hat{\mathbf{J}}^i, \mathbf{P}_o^i) < \tau) + \epsilon}, \quad (11)$$

where $d(\hat{\mathbf{J}}^i, \mathbf{P}_o^i)$ is the closest distance between hand joint $\hat{\mathbf{J}}^i$ and object surface $\mathbf{P}_o^i$ in i-th frame, τ is the contact threshold, $\mathbb{I}$ is the indicator function, and ϵ is a small constant to prevent division by zero.

*c) **Hand-Object Penetration Guidance.**:* We calculate the dot product between the hand-object offset and the object surface normal to measure the penetration as follows,

$$\mathcal{L}_{\text{penetration}} = -\sum_{i=1}^{l} \min(0, \tilde{\mathbf{F}}^i \cdot \mathbf{n}^i + \eta), \tag{12}$$

where $\tilde{\mathbf{F}}^i$ is the recalculated offset vector from the generated hand joint $\hat{\mathbf{J}}^i$ to the object surface $\mathbf{P}_o^i$ in i-th frame, $\mathbf{n}^i$ is the normal vector at the closest object point in i-th frame, and η is a small constant that defines the acceptable penetration threshold. The overall guidance loss function is

$$\mathcal{L}_g = \lambda_k \mathcal{L}_{\text{kinematic}} + \lambda_c \mathcal{L}_{\text{contact}} + \lambda_p \mathcal{L}_{\text{penetration}}. \tag{13}$$

The naive guidance strategy, which directly adds the computed gradient $\nabla_{\mathcal{H}^t}\mathcal{L}(\hat{\mathcal{H}}^0)$ to the sample, may misalign the final motions with the original data distribution, resulting in unnatural motions. Therefore, we aim to develop a guidance strategy that can achieve guidance targets while preserving the original distribution. Recently, a conditional diffusion model with the Spherical Gaussian constraint, named DSG [52] is proposed to offer larger, adaptive step sizes while preserving the original data distribution. However, its effectiveness has mainly been shown in image-generation tasks. Motivated by DSG, we introduce the Spherical Gaussian constraint to our sampling process, which ensures that the generated hand motion remains within the clean data distribution, resulting in natural and smooth hand motion generation.

We utilize closed-form solution $\mathcal{H}^{t*}$ that enforces sampling in gradient descent direction while preserving $\mathcal{H}^{t*}$ within noisy data distribution. This can ensure the generated $\mathcal{H}^0$ remaining within the clean data manifold:

$$\mathcal{H}^{t*} = \mu_t - \sqrt{d}\sigma_t \nabla_{\mathcal{H}^t}\mathcal{L}_g(\hat{\mathcal{H}}^0)/||\nabla_{\mathcal{H}^t}\mathcal{L}_g(\hat{\mathcal{H}}^0)||_2, \tag{14}$$

where σ_t represents the variance in time step t and d represents the data dimensions. To enhance generation diversity, we re-weight the direction of $\mathcal{H}^{t*}$ and the sampling point $\mathcal{H}^t$ similar to Classifier-free guidance:

$$\mathbf{D} = \mathcal{H}^t + w(\mathcal{H}^{t*} - \mathcal{H}^t), \tag{15}$$

$$\mathcal{H}^t = \mu_t + \sqrt{d}\sigma_t \mathbf{D}/||\mathbf{D}||, \tag{16}$$

where w represents the guidance rate that lies in [0,1] and $\mathcal{D}$ represents the weighted direction.

Since we represent the hand pose using hand joints, a post-optimization process is required to derive the corresponding MANO representation. We first calculate the global R by matching the actual MCP joint positions of the hand with its ideal position in the static state. The MCP joints are crucial for determining the global orientation of the hand because they serve as the primary pivot points for finger movement and are centrally located, providing a stable reference for the overall hand pose. By matching these joints, we can accurately capture the hand's global rotation, ensuring that the subsequent pose adjustments maintain anatomical correctness. We align the generated joints $\mathbf{J}$ with those from the MANO layer $\mathcal{M}(\cdot)$ via a joint loss $\mathcal{L}_{\text{joint}} = \|\mathbf{J} - \mathcal{M}(\mathbf{H})\|_2$. To preserve the natural and realistic movement of the hand, we apply pose angle constraints at the end of each iteration. These constraints limit the rotation range of hand joints across different axes, with strict controls specifically on the Y-axis rotation of the MCP joints and the Z-axis rotation of the other joints.

C. Motion Selection via Consistency Score

To ensure generated motions conform to gaze distributions, we evaluate gaze-interaction consistency from local and global perspectives (Fig. 3b). Locally, we compute a contact map $\mathbf{M}_c = \exp(-\alpha \cdot d(\mathbf{P}_h, \mathbf{P}_o))$ and a gaze map $\mathbf{M}_g = \exp(-\alpha \cdot d(\mathbf{G}, \mathbf{P}_o))$ based on distances from hand vertices $\mathbf{P}_h$ and gaze $\mathbf{G}$ to the object surface $\mathbf{P}_o$. The gaze-contact consistency score is $S_c = \frac{1}{|\mathbf{G}_\delta|}\sum_{i \in G_\delta} \mathbf{M}_{c_i}$, where $\mathbf{G}_\delta$ indexes object points within distance δ of $\mathbf{G}$. Globally, we quantify trajectory similarity using Dynamic Time Warping (DTW) [53]: $\text{DTW}(\mathbf{X}, \mathbf{Y}) = \min_\pi \sum_{(i,j)\in\pi} d(x_i, y_j)$, where π is the alignment path. The trajectory consistency score is defined as $S_t = \sum_{m \in \{l,r,o\}} \text{DTW}(\mathbf{G}, \hat{\mathbf{T}}_m)$. Finally, we select top-k motions $\langle \hat{\mathbf{H}}, \hat{\mathbf{O}} \rangle$ by maximizing the combined objective: $\arg\max_{(\hat{\mathbf{H}}, \hat{\mathbf{O}})}(S_c - \lambda_m S_t)$, where λ_m balances the two consistency terms.

V. Experiments

A. Data Split and Evaluation Metrics

We divided 1378 sequences into a training set with 1103 sequences and a test set comprising 275 sequences. To mitigate the model's potential overfitting to object-specific traits, the test set exclusively includes objects not represented in the training set. Our evaluation metrics includes three parts: 1) motion consistency with gaze and accuracy by evaluating the hand Mean Per Joint Position Error (MPJPE), object Mean Per Vertex Position Error (MPVPE), and the Final Object Location (FOL); 2) the stability and realism of hand-object interactions using the Contact Frame radio (CF) and the Penetration Depth (PD); 3) some basic generation metrics, including FID and Diversity. The FID quantifies the dissimilarity between the real and generated motions, while the diversity measures the variance of different motions. Additionally, we conduct a user study in which participants rated the gaze-motion consistency (GMC) and the interactions naturalness (IN) on a five-point scale. Higher scores reflect better motion generation quality.

B. Implementation Details

GHO-Diffusion is built upon a transformer encoder, following the approach of [50], with 1000 denoising steps. We utilize a single NVIDIA GeForce RTX 4090 GPU for training. To simplify the model, we randomly sample 500 points from the surface of the object. The batch size is set to 128, with a learning rate of 1e-5. The latent dimension is uniformly set to 128 for object diffusion and 256 for hand diffusion. For stage-1, the hyperparameters are set as$\lambda_\alpha = 10$, $\lambda_t = 30$, $\lambda_v = 10$, $\lambda_s = 1$. For stage-2, the hyperparameters are set as $\lambda_\beta = 100$, $\lambda_b = 100$. For HOI-Manifold guidance, we set $\lambda_k = 100$, $\lambda_c = 1000$, $\lambda_p = 100$ and guidance rate $w = 0.99$. We designed the hyperparameters based on our expertise to balance the weights of each loss component, ensuring they remain within appropriate magnitudes for optimal performance. In motion selection, we set $k = 1$ for quantitative experiments and user studies.

TABLE III
QUANTITATIVE COMPARISON BETWEEN BASELINES AND OUR METHOD. THE GROUND TRUTH DIVERSITY IS 4.227 AND DIVERSITY RESULTS ARE BETTER IF THE METRIC IS CLOSER TO THE REAL DISTRIBUTION.

Methods	Accuracy			Grasp Quality		Generation		Human Perception	
	MPVPE ↓	FOL ↓	MPJPE ↓	CF↑	PD ↓	FID ↓	Diversity ↑	GMC↑	IN↑
MDM*	146.8	228.0	166.6	62.55	2.52	0.071	4.174	3.70	3.40
Text2HOI*	254.2	286.7	182.2	63.98	2.57	0.183	4.134	3.05	3.75
OMOMO*	-	-	161.2	62.61	2.49	0.051	4.121	4.15	4.05
Ours	**117.4**	**146.5**	**144.8**	**68.10**	**2.46**	**0.044**	**4.217**	**4.75**	**4.60**

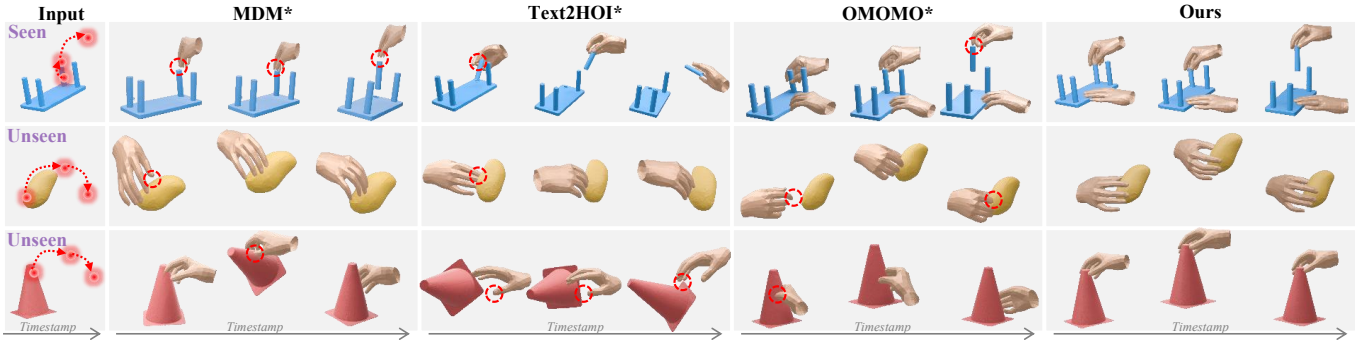


Fig. 4. Qualitative results between baseline methods and our method.

C. Baseline Design and Comparison

Lacking direct baselines for this novel task, we evaluate our approach against the following adapted methods.

1) Text2HOI:* The most closely related state-of-the-art work is Text2HOI [47], which generates a contact map conditioned on texts and then combines it to produce hand–object motions. Motivated by the relationship between gaze and contact, we adapt this approach by replacing the text encoder with our spatio-temporal gaze encoding to generate the contact map. Text2HOI employs a two-stage approach with a contact map as an intermediate representation to refine interaction modeling. However, its focus on single-frame spatial relationships and the use of modified position encoding overlook temporal coherence, which often leads to inconsistent hand–object interactions and noticeable jitter.

2) MDM:* We adapt MDM [50] by replacing its condition encoder with our spatio-temporal gaze encoding and introducing three auxiliary losses: $\mathcal{L}_{\text{obj}} = \sum_{i=1}^{L} \|\mathbf{T}_o^i - \hat{\mathbf{T}}_o^i\|_2$ and $\mathcal{L}_{\text{hand}} = \sum_{i=1}^{L} \|\mathbf{T}_h^i - \hat{\mathbf{T}}_h^i\|_2$ for independent translation constraints, and $\mathcal{L}_{\text{offset}} = \sum_{i=1}^{L} \|(\mathbf{T}_h^i - \mathbf{T}_o^i) - (\hat{\mathbf{T}}_h^i - \hat{\mathbf{T}}_o^i)\|_2$ to regularize the relative hand-object distance. The total training objective is $\mathcal{L}_{\text{MDM*}} = \lambda_\gamma \mathcal{L}_{\text{simple}} + \lambda_o \mathcal{L}_{\text{obj}} + \lambda_h \mathcal{L}_{\text{hand}} + \lambda_f \mathcal{L}_{\text{offset}}$, where we set $\{\lambda_\gamma, \lambda_o, \lambda_h, \lambda_f\} = \{20, 20, 20, 30\}$. MDM employs a simple classifier-free strategy for hand-object motion generation but lacks explicit constraints, resulting in inconsistent and physically implausible hand-object interactions.

3) OMOMO:* Object-Conditioned Hand Motion Synthesis is a key component of our method. Given the absence of a direct baseline for the full task, we adapt OMOMO [42] for comparison by utilizing our Stage-1 object motions as the required conditioning for its hand synthesis process. OMOMO generates wrist positions and optimizes subsequent trajectories based on the first contact frame to maintain interaction consistency and contact stability. However, this approach overly constrains motion diversity and lacks expressiveness, as wrist position alone is insufficient for fine-grained interaction modeling. In contrast, our method incorporates kinematic, contact, and penetration guidance for more realistic and dynamic interactions. As shown in Table III and Fig. 4, our approach significantly outperforms baselines. Note that MPVPE and FOL for OMOMO* are omitted (”-”) because it shares our stage-1 object results, making these metrics redundant for comparison.

D. Ablation Study

1) Results on Seen and Unseen Objects: To assess generalization, our test set (275 sequences) emphasizes novel scenarios: 215 feature entirely unseen objects, while 60 involve seen objects in complex tasks like assembly or chess. These scenarios rigorously test the model's robustness. As shown in Fig. 4 and Table IV, our method consistently produces natural, gaze-aligned motions, outperforming baselines across both seen and unseen categories. This demonstrates superior adaptability to varying interaction complexities without overfitting to object-specific traits.

2) Gaze Representation and Encoding: As the first method to use gaze as a condition for motion generation, we explore the effectiveness of different gaze representations and spatial-temporal feature encoding. The results are in Table V. We compare two gaze representations: gaze rays and gaze points. The encoding of gaze points has been detailed in the previous sections. In our framework, the gaze ray is formulated as a unit directional vector. To effectively incorporate this representation, we first process the gaze ray through a linear layer to obtain its embedding. Subsequently, we employ self-attention mechanisms to model its temporal dependencies. Finally, we integrate the gaze ray representation into the diffusion model using the same conditioning strategy as applied to gaze points. However, compared to gaze points, gaze rays provide only directional information and lack precise spatial relationships, weakening spatial modeling and ultimately affecting interaction quality. The ablation study on spatiotemporal encoding

TABLE IV
QUANTITATIVE COMPARISON ON **SEEN OBJECTS** AND **UNSEEN OBJECTS**.

Methods	Seen Objects							Unseen Objects						
	MPVPE ↓	FOL ↓	MPJPE ↓	CF↑	PD ↓	GMC↑	IN↑	MPVPE ↓	FOL ↓	MPJPE ↓	CF↑	PD ↓	GMC↑	IN↑
MDM*	147.5	214.4	165.8	42.30	2.17	4.05	3.55	146.6	231.8	166.8	68.19	2.62	3.40	3.35
Text2HOI*	247.5	275.1	179.0	54.66	1.98	3.25	4.25	256.1	289.9	183.1	66.58	2.73	3.05	3.75
OMOMO*	-	-	157.2	52.00	1.95	4.30	4.2	-	-	163.3	65.57	2.65	4.00	3.60
Ours	**113.1**	**133.3**	**146.6**	**59.28**	**1.92**	**4.90**	**4.75**	**118.7**	**150.2**	**144.3**	**70.55**	**2.61**	**4.70**	**4.55**

TABLE V
ABLATION ON GAZE REPRESENTATION AND ENCODING.

	MPVPE↓	FOL↓	GMC↑
gaze ray	126.7	189.2	3.95
w/o spatial	128.4	193.8	3.90
w/o temporal	119.3	163.4	4.15
w/o post-processing	134.8	179.2	3.95
Ours-	**117.9**	**157.6**	**4.25**

TABLE VI
ABLATION ON DIFFERENT GUIDANCE STRATEGIES.

	MPJPE ↓	CF↑	PD ↓	IN↑
w/o guidance	155.3	64.65	2.61	3.9
naïve guidance	156.8	63.66	**2.44**	3.5
Ours-	**154.3**	**65.77**	2.51	**4.45**

TABLE VII
ABLATION ON MOTION SELECTION.

	MPVPE↓	FOL↓	MPJPE↓	CF↑	PD↓	FID↓	GMC↑	IN↑
w/o select	117.9	157.6	154.3	65.77	2.51	0.048	4.25	4.45
w/o global	117.9	156.4	145.7	67.89	2.47	0.045	4.40	4.55
w/o local	117.5	148.9	149.8	65.96	2.50	0.046	4.35	4.45
Ours	**117.4**	**146.5**	**144.8**	**68.10**	**2.46**	**0.044**	**4.75**	**4.60**

highlights the importance of both temporal and spatial modeling for effective gaze representation. Temporal modeling filters out redundant and noisy data, allowing the model to focus on key gaze information, while spatial modeling captures the relationship between gaze and scene geometry, transforming gaze signals into meaningful interaction cues. The results demonstrate that our method effectively encodes both spatial and temporal features, whereas the absence of either component degrades encoding quality and negatively impacts first-stage performance. Furthermore, we conducted an ablation study on our gaze post-processing pipeline, which confirms its effectiveness in mitigating data noise and improving overall synthesis quality.

3) HOI-Manifold Guidance: Guidance-free methods suffer from contact inconsistencies and penetration errors. While naive guidance strategies can mitigate these issues, they tend to disrupt the learned diffusion process and introduce unnatural motion distortions. In contrast, our HOI-Manifold Guidance optimizes interactions while preserving the learned motion distribution, effectively striking a balance between physical constraints and motion naturalness. As shown in Table VI, this approach yields superior, physically plausible results.

4) Motion Selection via Gaze-interaction Consistency Score: Table VII validates that our consistency score selects motions better aligned with gaze cues. Specifically, local selection refines fine-grained hand-object contact details, significantly improving MPJPE, CF, and PD. Meanwhile, global selection enhances overall trajectory consistency with spatial gaze cues, leading to notable gains in MPVPE and FOL.

TABLE VIII
PERFORMANCE OF RECONSTRUCTION BASELINES.

Splits	Methods	MRRPE↓	MPJPE↓	SR↑
Val.	Arctic-SF [21]	111.2/43.5	**23.7**	20.3
	Arctic-LSTM [21]	114.9/45.2	24.0	22.5
	Kypt Trans. [54]	**108.9/42.6**	32.6	**28.4**
Test	Arctic-SF [21]	**65.1**/42.6	**22.8**	29.7
	Arctic-LSTM [21]	68.9/**42.4**	23.1	33.0
	Kypt Trans. [54]	111.4/42.5	27.6	**44.9**

VI. TASK: HAND-OBJECT MOTION RECONSTRUCTION

Leveraging rigorous data collection and precise cross-coordinate alignment, GazeHOI enables high-quality 3D hand-object reconstruction from visual cues. Given an interaction video and object mesh, the task aims to reconstruct continuous and natural hand-object motions for each frame.

Data and Metrics: We utilize 12 side views (3.0M images), split by subject: 8 for training, 1 for validation, and 1 for testing. Evaluation metrics include Mean Relative-Root Position Error (MRRPE) for relative root positions, local Mean Per Joint Position Error (MPJPE), and Success Rate [21] (percentage of predicted object vertices with L2 error $< 5\%$ of the object's diameter).

Baselines and Results: To establish a benchmark, we evaluate two baselines: (1) ArcticNet [21], including an image-based version (SF) and a video-based version (LSTM) for temporal reasoning; and (2) Keypoint Transformer [54], an image-based model effective for joint identification across challenging datasets like InterHand2.6M [55] and HO-3D [16]. Table VIII presents the quantitative results, validating GazeHOI's utility for reconstruction tasks beyond synthesis.

VII. CONCLUSION

In this paper, we introduced Gaze-guided Hand-Object Interaction Synthesis and GazeHOI, the first dataset capturing simultaneous 3D gaze, hand, and object interactions. We proposed GHO-Diffusion, a stacked diffusion model with HOI-Manifold Guidance to ensure physically plausible sampling. Furthermore, we leveraged gaze-contact and trajectory consistency to refine the motions. Experimental results validate the effectiveness of our method and the unique contributions of our dataset. Despite its diversity, GazeHOI currently excludes articulated objects, which present unique interaction challenges. Future work will extend this research to incorporate articulated interactions and multi-modal inputs, such as text and electromyography (EMG) data, to generate more context-aware interactive motions.

REFERENCES

[1] W. X. Schneider, "Vam: A neuro-cognitive model for visual attention control of segmentation, object recognition, and space-based motor action," *Visual Cognition*, vol. 2, no. 2-3, pp. 331–376, 1995. [Online]. Available: https://doi.org/10.1080/13506289508401737

[2] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[3] L. Desanghere and J. J. Marotta, "The influence of object shape and center of mass on grasp and gaze," *Frontiers in psychology*, vol. 6, p. 1537, 2015.

[4] A. A. Abdelrahman, T. Hempel, A. Khalifa, A. Al-Hamadi, and L. Dinges, "L2cs-net: Fine-grained gaze estimation in unconstrained environments," in *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*. IEEE, 2023, pp. 98–102.

[5] Y. Guan, Z. Chen, W. Zeng, Z. Cao, and Y. Xiao, "End-to-end video gaze estimation via capturing head-face-eye spatial-temporal interaction context," *IEEE Signal Processing Letters*, vol. 30, pp. 1687–1691, 2023.

[6] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz, "Weakly-supervised physically unconstrained gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9980–9989.

[7] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 334–352.

[8] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Improving driver gaze prediction with reinforced attention," *IEEE Transactions on Multimedia*, vol. 23, pp. 4198–4207, 2021.

[9] S. Zhang, Y. Tian, Y. Zhang, M. Tian, and Y. Huang, "Domain-consistent and uncertainty-aware network for generalizable gaze estimation," *IEEE Transactions on Multimedia*, vol. 26, pp. 6996–7011, 2024.

[10] J. Lou, H. Lin, P. Young, R. White, Z. Yang, S. C. Shelmerdine, D. Marshall, E. Spezi, M. Palombo, and H. Liu, "Predicting radiologists' gaze with computational saliency models in mammogram reading," *IEEE Transactions on Multimedia*, vol. 26, pp. 256–269, 2024.

[11] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 367–373, 2020.

[12] Y. Zheng, Y. Yang, K. Mo, J. Li, T. Yu, Y. Liu, C. K. Liu, and L. J. Guibas, "Gimo: Gaze-informed human motion prediction in context," in *European Conference on Computer Vision*. Springer, 2022, pp. 676–694.

[13] J. DelPreto, C. Liu, Y. Luo, M. Foshey, Y. Li, A. Torralba, W. Matusik, and D. Rus, "Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 800–13 813, 2022.

[14] F. Ragusa, A. Furnari, S. Livatino, and G. M. Farinella, "The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1569–1578.

[15] M. Eren Akbiyik, N. Savov, D. Pani Paudel, N. Popovic, C. Vater, O. Hilliges, L. Van Gool, and X. Wang, "G-memp: Gaze-enhanced multimodal ego-motion prediction in driving," *arXiv e-prints*, pp. arXiv–2312, 2023.

[16] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox, "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 9044–9053.

[18] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 138–10 148.

[19] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 20 953–20 962.

[20] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 21 013–21 022.

[21] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 12 943–12 954.

[22] Y. Liu, H. Yang, X. Si, L. Liu, Z. Li, Y. Zhang, Y. Liu, and L. Yi, "Taco: Benchmarking generalizable bimanual tool-action-object understanding," *arXiv preprint arXiv:2401.08399*, 2024.

[23] J. Zhang, H. Luo, H. Yang, X. Xu, Q. Wu, Y. Shi, J. Yu, L. Xu, and J. Wang, "Neuraldome: A neural modeling pipeline on multi-view human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 8834–8845.

[24] C. Zhao, J. Zhang, J. Du, Z. Shan, J. Wang, J. Yu, J. Wang, and L. Xu, "I'm hoi: Inertia-aware monocular capture of 3d human-object interactions," *arXiv preprint arXiv:2312.08869*, 2023.

[25] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu, "Oakink2: A dataset of bimanual hands-object manipulation in complex task completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 445–456.

[26] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019. [Online]. Available: http://arxiv.org/abs/1906.08172

[27] S.-Y. Chen, Y.-K. Lai, S. hong Xia, P. L. Rosin, and L. Gao, "3d face reconstruction and gaze tracking in the hmd for virtual interaction," *IEEE Transactions on Multimedia*, vol. 25, pp. 3166–3179, 2023.

[28] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri *et al.*, "Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 270–20 281.

[29] X. Pan, N. Charron, Y. Yang, S. Peters, T. Whelan, C. Kong, O. Parkhi, R. Newcombe, and Y. C. Ren, "Aria digital twin: A new benchmark dataset for egocentric 3d machine perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 133–20 143.

[30] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang, "Egobody: Human body shape and motion of interacting people from head-mounted devices," in *European Conference on Computer Vision*. Springer, 2022, pp. 180–200.

[31] M. Liu, S. Tang, Y. Li, and J. M. Rehg, "Forecasting human-object interaction: joint prediction of motor attention and actions in first person video," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 704–721.

[32] H. Yan, Z. Hu, S. Schmitt, and A. Bulling, "Gazemodiff: Gaze-guided diffusion model for stochastic human motion prediction," *arXiv preprint arXiv:2312.12090*, 2023.

[33] Z. Hu, S. Schmitt, D. Haeufle, and A. Bulling, "Gazemotion: Gaze-guided human motion forecasting," *arXiv preprint arXiv:2403.09885*, 2024.

[34] H. Li, X. Lin, Y. Zhou, X. Li, Y. Huo, J. Chen, and Q. Ye, "Contact2grasp: 3d grasp synthesis via hand-object contact constraint," *arXiv preprint arXiv:2210.09245*, 2022.

[35] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 107–11 116.

[36] S. Liu, Y. Zhou, J. Yang, S. Gupta, and S. Wang, "Contactgen: Generative contact modeling for grasp generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 609–20 620.

[37] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1471–1481.

[38] H. Zhang, Y. Ye, T. Shiratori, and T. Komura, "Manipnet: neural manipulation synthesis with a hand-object spatial representation," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–14, 2021.

[39] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2151–2162.

[40] O. Taheri, Y. Zhou, D. Tzionas, Y. Zhou, D. Ceylan, S. Pirk, and M. J. Black, "Grip: Generating interaction poses using spatial cues and latent consistency," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 933–943.

[41] K. Zhou, B. L. Bhatnagar, J. E. Lenssen, and G. Pons-Moll, "Gears: Local geometry-aware hand-object interaction synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 634–20 643.

[42] J. Li, J. Wu, and C. K. Liu, "Object motion guided human motion synthesis," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 6, pp. 1–11, 2023.

[43] J. Zheng, Q. Zheng, L. Fang, Y. Liu, and L. Yi, "Cams: Canonicalized manipulation spaces for category-level functional hand-object manipulation synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 585–594.

[44] S. Christen, M. Kocabas, E. Aksan, J. Hwangbo, J. Song, and O. Hilliges, "D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 577–20 586.

[45] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "Goal: Generating 4d whole-body motion for hand-object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 263–13 273.

[46] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "Imos: Intent-driven full-body motion synthesis for human-object interactions," 2023.

[47] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2hoi: Text-guided 3d motion generation for hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1577–1585.

[48] S. Christen, S. Hampali, F. Sener, E. Remelli, T. Hodan, E. Sauser, S. Ma, and B. Tekin, "Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions," *arXiv:2403.17827*, 2024.

[49] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, Nov. 2017.

[50] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[51] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[52] L. Yang, S. Ding, Y. Cai, J. Yu, J. Wang, and Y. Shi, "Guidance with spherical gaussian constraint for conditional diffusion," *arXiv preprint arXiv:2402.03201*, 2024.

[53] A. Seshan, "Using machine learning to augment dynamic time warping based signal classification," *arXiv preprint arXiv:2206.07200*, 2022.

[54] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 080–11 090.

[55] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer, 2020, pp. 548–564.