MAD-HOI: MASKED AUTOREGRESSIVE DIFFUSION FOR GENERATING ARTICULATED HAND OBJECT INTERACTIONS FROM TEXT

Anonymous authors
Paper under double-blind review

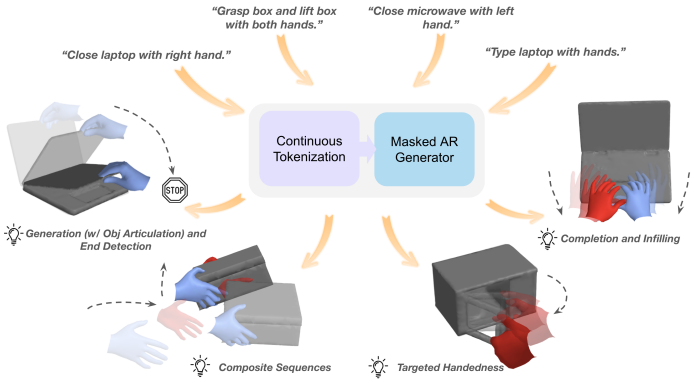


Figure 1: Beyond static grasps and atomic rigid-object motions, MAD-HOI supports articulated and composite interactions, motion completion and infilling, handedness control, and termination through end of motion detection.

ABSTRACT

Methods for text-based generation of hand–object interaction (HOI) sequences primarily focus on producing smooth, physically plausible trajectories. A truly utilitarian method should additionally support variable-length generation, composite motion sequences, motion completion and infilling, and reliable termination without compromising physical plausibility. Standard diffusion models for HOI generation are typically trained only for text-to-motion generation on atomic motions and require the motion length to be specified *a priori*. Autoregressive (AR) methods provide greater sequence-level flexibility, but commonly depend on discrete motion codes, which can lose contact-sensitive motion detail.

To address these key limitations, we present a model performing **M**asked **A**utoregression with **D**iffusion for **HOI** generation (**MAD-HOI**). Our method starts by encoding hand and object motions in a continuous latent space while keeping them disentangled to maintain stream-wise control. This is followed by a masked autoregressive transformer to predict context features that condition a flow-matching head. MAD-HOI is capable of motion generation for atomic and composite articulated sequences, conditioned motion completion and infilling, as well as EOM (End of Motion) prediction from a single training objective.

We provide comprehensive evaluations for these capabilities and benchmark our method on the ARCTIC and GRAB datasets. Our experiments show that MAD-HOI achieves the strongest text–motion alignment and distributional generation quality against baselines on both ARCTIC and GRAB, while maintaining competitive geometric accuracy and physical plausibility. Code will be made public.

1 INTRODUCTION

Hand manipulation is a fundamental human motor skills (Cheng et al., 2023), and offers a powerful lens into action understanding (Liu et al., 2024). The ability to predict its underlying motions is central to advancing both robotic manipulation (Yang et al., 2025; Singh et al., 2025) and realism in virtual reality (Mangalam et al., 2024). Studying hand movements for object manipulation through the lens of 3D computer vision is commonly referred to as Hand Object Interaction (HOI). In this

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

paper, we focus on a specific HOI task - Generation of Hand and Object Motions from Text inputs. In particular, we assume that object geometry and its articulating segments are known, and that the intent of the interaction is provided via language instruction.

Diffusion-based (Sohl-Dickstein et al., 2015) HOI generation methods Text2HOI (Cha et al., 2024), DiffH2O (Christen et al., 2024), LatentHOI(Li et al., 2025a) produce smooth, realistic motions but are structurally 'atomic': the absence of a strong long-horizon motion prior restricts these approaches to short indivisible movement primitives. Moreover, sequence lengths must be provided by a separate predictor (Text2HOI) or by the user (DiffH2O, LatentHOI). With these methods, generating longer composite sequences from chained instructions is undefined, and trajectories for both hands are generated at all times, with the unused hand masked post-hoc for single-hand prompts.

A complementary line of work - Jiang et al. (2023); Huang et al. (2025) - offers exactly the missing controls, but pays for them in motion quality. These methods use vector quantization (VQ) to represent motion as discrete tokens and introduce these tokens to an LLM as a 'new language'. They gain variable-length generation, completion, and composition through next-token prediction, and termination from the standard EOS (End of Sequence) token. However, quantizing contact-critical motion is destructive: in our tokenizer comparison study (section 3), reconstruction from discrete tokens through a VQ-VAE has much higher penetration and lower contact than with continuous latents from our VAE, and produces visibly spikier jerk profiles (see supplementary). The information lost at tokenization caps the quality of everything generated downstream.

We address this trade-off with a framework that retains the compositional and variable-length benefits of AR modeling while maintaining a continuous motion representation. Our method employs the **M**asked **A**utoregressive **D**iffusion paradigm for text to **HOI** sequence generation, and is hence named **MAD-HOI**. We make certain design choices to enable additional capabilities and maximize the expressivity and text-alignment of generated motions.

We begin by encoding the object, left-hand, and right-hand motions into disentangled continuous latent streams through a Variational Autoencoder (VAE). The VAE is cascaded to condition hand motions on object trajectories and is supervised with contact-aware geometric losses. We then train a bidirectional masked autoregressive transformer to predict conditioning signals for a flow-matching head that generates motion latents. This generator is designed to respect the individual motion streams as separate modalities and predict EOM tokens. Through a single training objective of masked conditional flow matching, and without task-specific heads or post-processing, MAD-HOI learns to respect handedness in text and generate the respective motion streams, complete and infill motion sequences smoothly, generate longer-horizon composite motions and predict EOM tokens eliminating the need for user-specified sequence lengths, as seen in (table 1).

Our contributions are summarized as follows:

- We demonstrate that continuous motion latents yield smoother reconstructions than discrete VQ tokens, and propose a cascaded VAE that encodes short temporal sequences into a structured and disentangled latent space.
- We show that combining our latent structure with a Masked Autoregressive Diffusion model provides a set of capabilities no existing diffusion HOI method has: varying-length composite generation with learned EOM-based sequence termination, conditioned completion, and infilling while maintaining targeted handedness with genuinely idle hands.
- We present comprehensive evaluations and further evaluate text-motion alignment with a protocol wherein the motion evaluator encodes object motion and articulation alongside hand motion, unlike prior protocols that consider hand motion only.

2 Related Works

Hand–object interaction generation. In full-body HOI, CG-HOI (Diller & Dai, 2024) jointly diffuses contact with human and object motions; OOD-HOI (Zhang et al., 2024b) uses reciprocal diffusion and contact-guided refinement for unseen-object and action generalization; and GenHOI (Li et al., 2025b) reconstructs sparse interaction keyframes before synthesizing temporally coherent 4D sequences with contact-aware diffusion. ARDHOI (Geng et al., 2025) combines continuous HOI representations with causal autoregressive diffusion. However, its next-token formulation is naturally suited to forward generation, whereas masked bidirectional prediction better suits completion and infilling. These methods lack object and per-hand stream separation for modality control.

Diffusion-based methods for fine-grained hands-only HOI generation typically involve multiple stages, limited sequence-level controllability, and externally supplied motion lengths. Text2HOI (Cha et al., 2024) first predicts contact maps and then generates motion conditioned on them through diffusion, followed by a penetration-reduction refiner. DiffH2O (Christen et al., 2024) separates grasping and interaction into two diffusion stages and represents motion in an object-centered coordinate frame. LatentHOI (Li et al., 2025a) similarly adopts an object-centered representation and factorizes generation into latent wrist trajectories and hand-pose decoding. These decompositions can aid unseen-object generalization, but introduce specialized stages and primarily target atomic interactions whose duration must be supplied or separately predicted. They also generate both hand trajectories irrespective of handedness, requiring the unused hand to be suppressed after generation rather than modeled as part of the output distribution. OpenHOI (Zhang et al., 2025b) extends a Text2HOI-style generator with multimodal affordance reasoning, physics-based refinement, and a separate post-hoc module for compositing atomic motions. Autoregressive approaches provide stronger sequence-level flexibility. HOIGPT (Huang et al., 2025) discretizes hand and object trajectories into VQ-VAE codes and trains a language model over motion and text tokens, naturally supporting variable-length generation, completion, and compositional sequencing. Its capabilities, however, depend on quantized motion representations, which can discard fine-grained kinematic and contact information. MAD-HOI instead retains continuous motion latents while explicitly structuring the sequence into object, left-hand, and right-hand streams, allowing the model to generate asymmetric bimanual interactions, represent inactive hands directly, and predict the termination of each modality without post-hoc masking or an external length predictor.

Concurrent work, HO-Flow (Chen et al., 2026) independently combines continuous temporal latents with masked autoregressive latent diffusion. The method learns an interaction-aware VAE that projects the object point cloud into the local coordinate systems of the MANO kinematic chain, capturing hand-relative geometry alongside global hand and object motion. It then concatenates the resulting hand–object latents into unified temporal tokens and trains a generator similar to ours. While HO-Flow reports improved physical plausibility over conventional diffusion-based baselines, it does not demonstrate capabilities beyond input-length-constrained text-to-motion generation. Moreover, its validation misses out on articulated-object interactions, text–motion alignment and standard distributional metrics. We omit this baseline as the code is not publicly available and their reported results are not directly comparable as we use different datasets and preprocessing protocols—including different processing of GRAB (Taheri et al., 2020).

Capability	Text2HOI	DiffH2O	LatentHOI	OpenHOI	HOIGPT	MAD-HOI
Continuous Repr.	✓	✓	✓	✓	✗	✓
Motion Generation	✓	✓	✓	✓	✓	✓
Auto-termination (EOM)	✗	✗	✗	✗	✓	✓
Motion Completion	✗	✓	✓	✗	✓	✓
Motion Infilling	✗	✗	✗	✗	✓	✓
Composite Motion	✗	✗	✗	✓	✓	✓
Captioning	✗	✗	✗	✗	✓	✗

Table 1: **Capability comparison** of MAD-HOI with relevant baselines (determined from their papers and public code.)

3 Tokenization: Discrete vs Continuous

Autoregressive modeling operates on a tokenized latent space, and the tokenizer bounds the quality of downstream generations. VQ-VAE-based methods Jiang et al. (2023); Huang et al. (2025) encode motion into discrete codebooks and train an LLM to predict the code indices; we argue that discretization loses information HOI cannot afford to lose, as small quantization errors in wrist pose or finger articulation translate directly into penetration, broken contact and jitter.

To demonstrate this, we train our VAE and an equivalent VQ-VAE architecture (similar to Huang et al. (2025)) and compare their reconstructions, measuring penetration and contact percentages following Zhang et al. (2025a) and, following Cho et al. (2025), jerk profiles (the third derivative of position) which exposes high-frequency jitter that mean errors hide; plots presented in the Appendix A.2. ***The VAE reconstructs with penetration in only 2.1% of frames and contact in 92.2% frames, against 24.7% and 77.9% of the frames respectively for the VQ-VAE. And the jerk profiles show that the VAE tracks the ground truth closely and produces less jerky (jittery) motion, especially when it comes to reconstructing the longer tail of static hands.*** This motivates continuous motion latents over discrete motion tokens for autoregressive HOI generation.

4 METHOD

MAD-HOI first uses a cascaded VAE (left in fig. 2) to create motion latents for object and hand trajectories. These motion latents then undergo masking before being passed to the autoregressive diffusion generator of our framework (right in fig. 2), which consists of a masked autoregressive transformer producing conditions for the diffusion head at masked positions. Our method derives its capabilities from two streams - 1. The bidirectional autoregressive model with iterative masked denoising allows the generator to look at the full unmasked temporal context while generating each motion latent, giving rise to conditioned completion and infilling (depending on the masking). 2. Our VAE with it's disentangled the latent structure enables modality control and stream-wise generation/completion/infilling which improves 'handedness'. It further learns 'idle hand' latents which enable the generator to get the right context for predicting EOM and therefore, sequence termination.

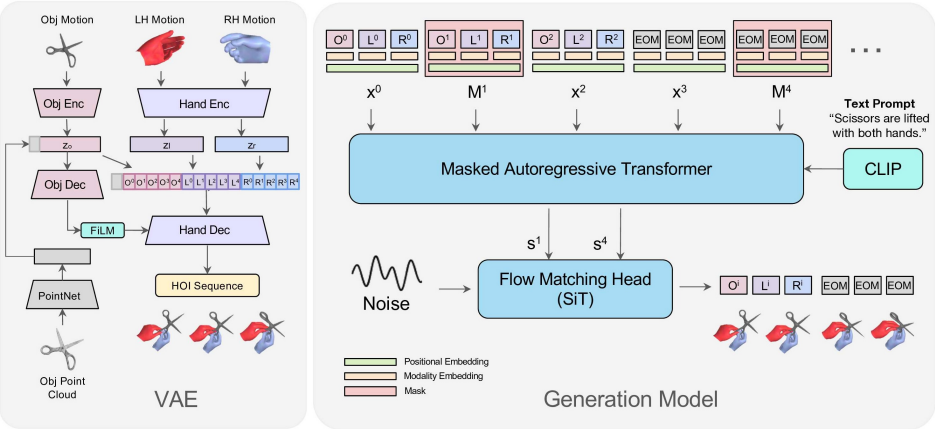


Figure 2: **MAD-HOI architecture.** Left: Our cascaded VAE builds continuous motion latents for object and hand trajectories. Right: Our generation model consists of a masked autoregressive transformer that predicts conditioning signals for a flow-matching head - generating motion latents for masked tokens, enabling smooth, text-based HOI generation, completion, infilling and termination.

Data: We focus on motion datasets that capture one-handed and bimanual object interactions in 4D. GRAB (Taheri et al., 2020) captures full-body sequences with rigid objects, of which we use only the hand interaction data; ARCTIC (Fan et al., 2023) consists of single-joint articulating objects. With the preprocessing protocol and text annotations from (Cha et al., 2024), GRAB has 1.6k motion sequences for 51 objects and ARCTIC 4.6k atomic sequences spanning 11 objects; this difference in trajectories per object prompts us to maintain separate experiments for the two datasets.

The object trajectory is a list of global poses $O = (T_o^t, R_o^t, \alpha_o^t)$, where T_o is the normalized global 3D location of the object center, R_o its 6D rotation, α_o an optional articulation value for articulated objects, and t the time frame. Hand poses are based on MANO (Romero et al., 2017): each hand's trajectory is $H = (\tau_h^t, \phi_h^t)$, with τ_h the normalized global 3D wrist position and ϕ_h the 6D rotations of the wrist and the other 15 MANO joints, plus a binary left/right identifier appended separately.

4.1 VARIATIONAL AUTOENCODER

To enable modality control in generation, our tokenizer must keep the three motion modalities (object, left hand and right hand) as separate latent streams. Following prior works Jiang et al. (2023); Meng et al. (2025); Huang et al. (2025), we use 1D conv + resnet blocks in the encoders and decoders of our cascaded VAE. We encode the object motion and the hand motions through different branches. We further recognize that hand motions depend heavily on the object motion and therefore design a cascaded network structure that decodes hands conditioned on the decoded object motion. **Object branch:** The object encoder encodes O into a 512-dimensional latent z_o, and the object decoder receives pointnet features from a canonical object point cloud, which are concatenated with z_o to reconstruct object trajectories. **Hand branch:** The hand encoder and decoder maintain one set of weights to separately process two streams of input, H_{left} and H_{right}, latents z_l and z_r for the left and right hands respectively. This enables co-learning and efficiently embeds both hand motion streams into a shared latent space while keeping them differentiated through an identifier value. The object latent z_o is concatenated with z_l and z_r and passed to the hand decoder to jointly decode the

full HOI sequence. The hand decoder also applies the reconstructed O as conditions through a FiLM layer (Perez et al., 2018).

In addition to the motion reconstruction objective, we force our VAE to reconstruct static poses over short horizons as 'idle latents'. We do this to have hand inactivity represented by static (or 'idle') poses, lending to true 'handedness' without post-hoc masking of either hand. See section 4.2 for the use. Each encoder predicts a posterior mean and variance, and in the generator training, motion latents are drawn with the reparameterization trick, $z = \mu + \sigma\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. A KL-divergence loss regularizes each posterior toward a unit Gaussian. This regularization matters downstream: the flow-matching head of our generator regresses velocities toward samples from this latent space, and a bounded, smoothly populated space is a more stable regression target than the unconstrained latents of a deterministic autoencoder (see supplementary for this comparison).

We use a window length of 152 frames with a 4-frames-to-1-latent temporal compression throughout, giving at most 38 latents per modality (see fig. 2 for the architecture). The latents are 512 dimensional. The VAE is trained to reconstruct HOI sequences with L1 reconstruction loss and three geometric losses (definitions in the supplementary): a contact loss $\mathcal{L}_C$ penalizing hand joints that stray beyond a threshold distance ϕ from the object surface, a penetration loss $\mathcal{L}_{pen}$ penalizing hand vertices inside the object mesh, and a distance map loss $\mathcal{L}_{DM}$ (Liang et al., 2024) aligning predicted and ground-truth joint-to-object distance maps in near-contact regions. The total loss combines L1 losses over the translation and rotation components of the object and hand trajectories with the geometric losses and the KL term. Weights are detailed in the Appendix A.3.4 and network specification in Appendix A.4.

$$\mathcal{L} = \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_C + \gamma\mathcal{L}_{pen} + \delta\mathcal{L}_{DM} + \lambda\mathcal{L}_{KL} \tag{1}$$

4.2 Autoregressive Diffusion

Unlike diffusion models, autoregressive models factorize the conditional likelihood of a sequence, $p(x^{1:n} \mid c)$, into chained conditionals $p(x^i \mid x^{<i}, c)$, generating each token conditioned on the previously generated tokens and the input condition (c). Extending diffusion models for AR prediction with an MSE objective over ϵ_θ or x_0 collapses the generation task into simple regression with gaussian probability, failing to model the chained conditional probabilities that define $p(x^{1:n}|c)$ in true autoregressive generation. Recent advances in image generation (Li et al., 2024) highlight the promise of continuous autoregressive generation, where logits from an AR model are used as conditioning signals for a continuous sampling network to better approximate underlying data likelihoods with chained probability.

At each autoregressive iteration in MAD-HOI, interaction sequences are encoded into motion latents through the VAE and masked through a cosine schedule (Li et al., 2024). The unmasked latents form the visible context $\mathbf{um}$, while the masked latents $\mathbf{m}$ represent the 'tokens' that must be generated under condition c. Generation proceeds over K unmasking iterations, with $\mathbf{m}_j$ denoting the subset of masked latents generated at iteration j; see eq. (2)

$$p(x'_{1:n} \mid c) = \prod_{j=1}^{K} p(\mathbf{m}_j \mid \mathbf{um},\, \mathbf{m}_{<j},\, c)\,. \tag{2}$$

4.2.1 Masked Autoregressive Transformer

The transformer is tasked with capturing the context of time-variant motion data, and provide meaningful conditions s for the flow matching head. In our case, an interaction sequence encoded with our VAE results in three motion latents, z_o, z_l, z_r which are all considered as different motion modalities in our framework. We refer to this triplet of latents as x and maintain this structure per timeframe by interleaving latents from the three modalities and attaching one positional encoding to each triplet. The latent from each modality also receives its own learnable ***'modality embedding'***. Since our VAE has a decoupled latent space for the hands, when the generation branch recognizes that a hand token corresponds to an 'idle' hand, it is able to predict latents corresponding to 'idleness' leading to the ability to understand 'hand use'. While Text2HOI, DiffH2O and LatentHOI generate trajectories for both hands irrespective of the single-hand text prompt and mask the unused hand for visualizations, our method actually predicts static poses for unused hands - leading to ***'handedness'.*** See Appendix A.7.2 for validation of handedness and idle or unused hands. Further, we pad all non-valid frames within our input with learnable **End of Motion** $\langle EOM \rangle$ **tokens**, while continuing the

interleaved structure. The model learns to associate this token with the long tail of 'idle' latents and predict the end of generation - meaning that we do not require an input sequence length for inference. A sample input sequence to the generator is provided in the Appendix A.3.2.

Given the unmasked motion latent sequence $\mathbf{um}$ as previously generated motion latents, the masked autoregressive transformer g produces conditions $s = g(\mathbf{um})$ for the diffusion branch to generate latents at masked positions m. Our transformer uses bidirectional attention (Meng et al., 2025) and a single transformer layer (Vaswani et al., 2017) with Adaptive Layer Norm (Peebles & Xie, 2023). The text caption is fed into the CLIP encoder and integrated into the transformer's latent space through a linear projection.

4.2.2 DIFFUSION:

The output s from the Masked Autoregressive Transformer is fed into the flow matching head as conditions. We call this per-token sampler the diffusion head for continuity with Li et al. (2024); Meng et al. (2025), though its training objective is flow matching. This network produces the motion latent x'^i at each masked position i by iterating the reverse step $x'^i_{t-1} \sim p(x'^i_{t-1} \mid x'^i_t, t, s_i)$ over timesteps t.

During the diffusion head training, k random motion latents are masked with a learnable mask vector following the cosine masking schedule from MoMask Guo et al. (2024). The autoregressive model learns to provide accurate signals based on the unmasked latents given the timestep and text condition, and the diffusion head uses this signal to predict a velocity $v(x, t)$. Following SiT Ma et al. (2024), we corrupt the ground-truth latent x_0 along the linear interpolant $x_t = \alpha_t x_0 + \sigma_t \epsilon = (1-t)x_0 + t\epsilon$, where ϵ is gaussian noise, t is the continuous timestep, and $\dot{\alpha}_t, \dot{\sigma}_t$ denote the time derivatives of the interpolation coefficients (the full velocity-field definition is provided in the Appendix A.3.3). We apply the following loss over velocity prediction:

$$\mathcal{L}_{\text{GB}} = \int_0^T \mathbb{E}_{\mathbf{v},t}\Big[\big\|\mathbf{v}_\theta(\mathbf{x'}^i_t \mid t, g(\mathbf{u}_m)) - \dot{\alpha}_t \mathbf{x'}^i_0 - \dot{\sigma}_t \boldsymbol{\epsilon}\big\|^2\Big]\, \mathrm{d}t, \tag{3}$$

At inference, given previous motion latents $\mathbf{um}$, mask vectors are added to the latent sequence, allowing the autoregressive model to generate signals for masked positions. The velocity prediction uses ODE sampling over 18 refinement steps (re-masking), taking Euler steps $\mathbf{x}^i_{t-1} = \mathbf{x}^i_t + \Delta t \cdot \mathbf{v}_\theta(\mathbf{x}^i_t \mid t, s^i)$ of size Δt. The generator architecture is a transformer with single attention block and 16 heads, followed by the SiT-XL flow matching head. All training details are in Appendix A.3.4 and network specification is in Appendix A.4.

4.3 BASELINES, METRICS, AND EVALUATION PROTOCOL

We consider recent open-sourced methods addressing the text-to-motion objective as baselines; some relevant contemporaneous approaches are omitted due to lack of public code. Our data preprocessing mirrors the Text2HOI protocol, making their available checkpoints valid for our evaluations. DiffH2O, assumes a grasping+interaction split, and uses the MANO PCA representation with an-object centric coordinate system (as does LatentHOI) leading to motions relative to the first frame. We therefore preprocess our data according to match their representations, but maintain our global coordinate system, and train both methods from scratch. OpenHOI inherits Text2HOI's diffusion generator, however it needs to be retrained due to mismatch in scales of supervision.

Metrics: ***For generation***, we report Top-3 Retrieval Accuracy (R@3), FID, KID, Diversity, and Matching Score, all computed in the embedding space of our evaluator over 5 repeats; definitions follow standard practice (Guo et al., 2022a) and are detailed in the appendix. ***For geometric accuracy***, we compute Accel - mean per-vertex acceleration magnitude over the MANO mesh (scaled by 100); Penetration % - the percentage of frames where any hand vertex penetrates the object by more than 1 cm; Contact % - the percentage of frames with a hand vertex within 5 mm of the object surface; Net Contact - a tradeoff metric computed as $C * (1 - P)$ to reward contact without penetration. We additionally report the fine-grained physical-plausibility suite of metrics from LatentHOI (Li et al., 2025a), all evaluated per active hand and averaged over the frames in which that hand is in contact with the object. Interpenetration Volume (IV, cm^3) is the volume of hand–object overlap, obtained by voxelizing the object at a 5 mm pitch and counting voxel centers that fall inside the hand mesh (a thin-shell object convention). Interpenetration Depth (ID, cm) is the deepest single hand-vertex penetration into the object. Contact Ratio (CR, %) is the fraction of hand vertices lying within 5 mm of the object surface. Interpenetration-Volume per Unit contact area (IVU, cm^3/cm^2) divides the total interpenetration volume by the total contact area, expressing penetration relative to

Dataset	Method	Acc.Top-3 ↑	FID ↓	KID ↓	Diversity →	Matching Score ↓
ARCTIC	*Ground Truth*	0.991	–	–	20.235	1.295
	Text2HOI	0.926	0.528	0.008	20.110	2.578
	DiffH2O	0.715	1.138	0.025	15.246	9.791
	LatentHOI	0.941	1.053	0.009	20.364	3.325
	OpenHOI	0.946	0.495	0.010	20.159	2.292
	MAD-HOI (Ours)	**0.991**	**0.071**	**0.004**	**20.312**	**1.353**
ARCTIC (Ablation)	w/o EOM Pred.	0.928	1.132	0.010	19.467	2.944
	w/o Mod. Tokens	0.542	56.736	0.787	17.641	11.993
ARCTIC Composite	*Ground Truth*	0.890	–	–	20.118	3.320
	MAD-HOI (Ours)	0.869	0.142	0.002	20.450	3.693
GRAB	*Ground Truth*	0.939	–	–	12.871	3.188
	Text2HOI	0.745	3.661	0.088	**13.177**	4.941
	DiffH2O	0.838	1.923	0.029	12.250	4.746
	LatentHOI	0.797	5.616	0.146	13.220	4.698
	OpenHOI	0.677	4.502	0.131	12.316	5.429
	MAD-HOI (Ours)	**0.932**	**1.077**	**0.014**	13.217	**4.041**
GRAB (Ablation)	w/o EOM Pred.	0.930	1.471	0.008	14.321	2.923
	w/o Mod. Tokens	0.737	11.024	0.292	11.330	5.772

Table 2: Retrieval and generation metrics on ARCTIC and GRAB. ↓ indicates lower is better, ↑ indicates higher is better, and → indicates closer to ground truth is better. Best result per metric is in **bold**.

Dataset	Method	Accel↓	Pen.%↓	Con.%↑	NC↑	IV↓	ID↓	CR↑	IVU↓	Phy↑
ARCTIC	*Ground Truth*	0.18	7.86	80.77	74.42	7.41	0.85	13.42	0.13	98.67
	Text2HOI	0.19	22.74	**92.90**	71.77	7.55	0.94	14.32	0.15	88.99
	DiffH2O	0.64	19.98	91.64	73.33	8.99	1.03	13.05	0.16	**93.13**
	LatentHOI	0.31	**12.37**	65.30	57.23	11.24	1.15	12.59	0.18	60.25
	OpenHOI	0.40	20.52	73.58	58.48	9.77	1.05	13.48	0.17	77.90
	MAD-HOI (Ours)	**0.18**	<u>12.89</u>	<u>84.47</u>	**73.58**	**7.52**	**0.94**	**14.76**	**0.14**	<u>92.65</u>
ARCTIC (Ablation)	w/o EOM Pred.	0.23	21.01	73.71	58.22	9.37	1.10	13.07	0.17	80.20
	w/o Mod. Tokens	0.44	11.20	49.50	43.96	7.96	1.32	11.06	0.177	58.4
ARCTIC Composite	*Ground Truth*	0.17	7.0	80.5	74.86	8.18	0.83	14.38	0.136	98.8
	MAD-HOI (Ours)	0.20	17.6	81.0	66.74	9.38	1.04	14.87	0.163	87.6
GRAB	*Ground Truth*	0.31	30.64	95.91	66.52	5.17	2.16	9.45	0.14	99.24
	Text2HOI	0.47	45.11	**97.86**	53.72	9.97	2.35	**13.59**	0.19	**94.49**
	DiffH2O	0.49	**31.34**	<u>92.71</u>	**63.66**	8.15	2.51	10.41	0.16	<u>93.20</u>
	LatentHOI	0.74	35.45	85.35	55.09	11.37	2.49	11.26	0.21	79.55
	OpenHOI	0.88	47.11	92.28	48.80	10.76	2.50	12.41	0.20	86.60
	MAD-HOI (Ours)	**0.38**	<u>34.60</u>	92.60	<u>60.56</u>	**7.98**	**2.07**	<u>12.42</u>	**0.15**	87.01
GRAB (Ablation)	w/o EOM Pred.	0.32	35.6	90.2	58.08	7.5	2.14	11.56	0.17	83.3
	w/o Mod. Tokens	0.59	24.9	63.0	47.31	8.45	2.59	9.15	0.21	56.4

Table 3: Geometric and physical-plausibility metrics (defined in section 4.3). ↓ lower / ↑ higher is better; best non-GT value per column in **bold**, second-best <u>underlined</u>.

how much of the hand actually contacts the object. And finally, Physical plausibility (Phy, %) is the fraction of object-moving frames (mean per-vertex object displacement > 1 cm between consecutive frames) in which the hand is simultaneously in contact, capturing whether the hand drives object motion rather than the object drifting on its own. ***For completion and infilling***, we report Average and Final Displacement Errors ADE and FDE (Yuan & Kitani, 2020). ***For EOM prediction***, the Mean and Median AE in no. of frames (with 4-frame and 8-frame tolerance variants, and and interval %) from ground truth sequence lengths. All metric definitions are in Appendix A.6.

Evaluation Protocol: For our generation metrics, we adopt the feature-based evaluation protocol from Guo et al. (2022a), which learns a shared text–motion embedding space using separate text and motion encoders. The two branches are trained with a margin-based contrastive objective that pulls matched text–motion pairs together and pushes mismatched pairs apart. All prior works using this or a similar protocol adopt the full-body motion encoder for the hand trajectories only, ignoring object motion. Since the object trajectory and articulation are key to the 'action' specified in the text, we train the evaluation pipeline from scratch to encode motion latents for the hands and the object jointly, before minimizing distances between motion features and their positive text pairs. More details in Appendix A.5. MAD-HOI and all baselines are evaluated uniformly with this protocol.

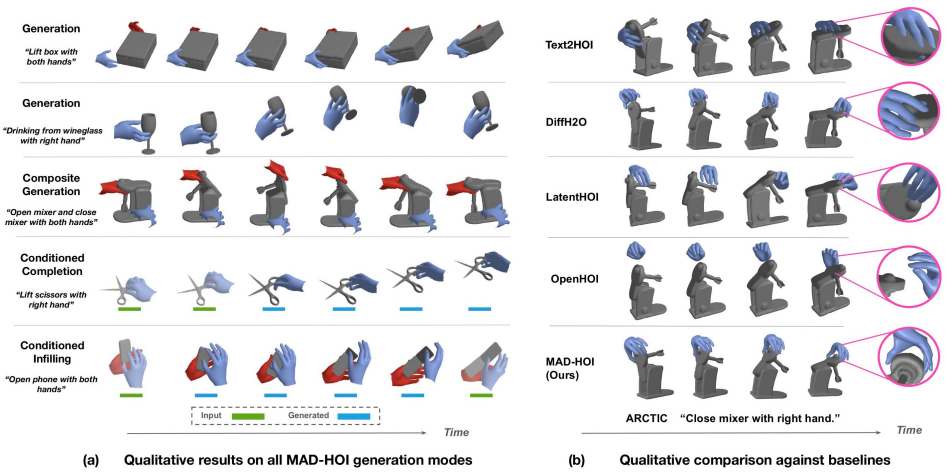


Figure 3: (a) Qualitative results showcasing all capabilities of MAD-HOI. (b) Comparison of grasp and motion against baselines.

5 RESULTS

Text-to-Motion Generation: We provide qualitative results for all our capabilities and visual comparisons against baselines in fig. 3 (GRAB comparisons in fig. 6). Quantitative comparisons with baselines on both metric families are presented in table 2 and table 3. MAD-HOI achieves the best retrieval accuracy, FID, KID, and Matching Score on both the datasets. While it achieves the highest Diversity for ARCTIC, it trails by a small margin for GRAB, see table 2. OpenHOI performs the second best on ARCTIC while DiffH2O is the second best method on GRAB. MAD-HOI's generations show the most gains in text-alignment metrics like FID and Matching Score.

When it comes to geometric metrics (table 3), we note that MAD-HOI performs well on ARCTIC. It has the best performance on 6 out of the 9 metrics, coming in second on Physical Plausibility, Penetration and Contact %s. We realize that since contact and penetration exist in a tradeoff, we compute NC to show a balance between the two and here MAD-HOI has the highest NC value. The picture is more blurry for GRAB though. MAD-HOI still has the smoothest generated motions (with the lowest acceleration and closer to ground truth (GT) as with ARCTIC). However, it trails in Contact % and CR. We still outperform the baselines in IV, ID and IVU and clock in the second best values for Penetration %, NC. However, the lower physical plausibility points to loss of contact when the object is off the ground plane and moving fast - something more characteristic of the GRAB dataset where the object trajectories span large areas. Overall, MAD-HOI is the strongest method for text-aligned generation with competitive geometric metrics. No baseline method outperforms MAD-HOI across both metric families consistently. It is also the preferred generation method in our user study discussed in Appendix A.7.1.

Composite Motion Generation: The MAD-HOI generator is trained with bidirectional attention and is pushed to predict EOM tokens. This coupled with iterative masked denoising at inference, gives MAD-HOI the ability to generate motions of varying lengths within its max context window, meaning that it can generate composite motions based on chained prompts up to 152 frames. We train the MAD-HOI generator on a bigger ARCTIC training dataset where we include 30% of the 2197 compound motions (up to 4 atomic motions chained). This debiases EOM likelihoods from atomic motion lengths and introduces words like 'and' to the generator. Many compound combinations are unseen in validation and metrics from table 2 and table 3 show that MAD-HOI has minimal quality degradation in composite generations. We omit this experiment on GRAB: its sequences are longer leaving very few valid testing sequences below the maximum context window.

Conditioned Completion and Infilling: During inference, masked out tokens beyond a few starting frames in the sequence result in completion, and in the center, result in infilling. Our generator is context-aware, its latent prediction is conditioned on whatever unmasked input is provided. We therefore report the ADE and FDE for masked frames with the closest semantic GT samples rather than raw trajectory errors in meters. In (section A.7.4), errors largely decrease monotonically as more ground-truth latents are provided to the generator as context, as is expected. At 2 conditioning latents, completion quality is close to atomic generation while infilling is harder, due to the stricter

constraint of matching context on both sides. Our HOI baselines do not perform completion/infilling, we therefore include simple protocol-matched kinematic baselines in Appendix A.7.4.

Sequence Termination with EOM: When the MAD-HOI generator predicts latents close to the learned EOM token for the masked tokens of all 3 modalities, the end of the sequence is detected and the trailing frames are discarded as the sequence is terminated. No separate length predictor needs to be trained or queried. We compare our implicit sequence termination performance indirectly by comparing against the dedicated cVAE length regressor of Text2HOI in table 5 which is explicitly trained with supervision on sequence lengths. The specialist regressor is more accurate on raw error, as expected. A single prompt, however, maps to many valid durations, so we also evaluate against the $[q_{10}, q_{90}]$ interval of demonstrated lengths per condition: MAD-HOI's predicted endpoints fall inside this valid range for 78.4% (ARCTIC) and 70.5% (GRAB) of prompts, and miss it by only 2–5 frames on average. Termination thus comes at no extra cost to the single generation objective or number of parameters, with modest accuracy against a specialist module. Appendix A.7.3 outlines why learned EOM is better than user-specified EOM.

		ADE ↓					FDE ↓					FID ↓	Diversity ↑
Task Tokens (#):		1	2	3	4	5	1	2	3	4	5	2	2
Completion	ARCTIC	0.649	0.617	0.613	0.608	0.592	0.816	0.787	0.789	0.737	0.733	0.173	20.325
	GRAB	0.748	0.776	0.738	0.749	0.724	0.963	1.019	1.001	0.960	0.882	1.719	13.010
Infilling	ARCTIC	0.783	0.707	0.675	0.639	0.629	0.833	0.766	0.719	0.691	0.684	0.182	19.2136
	GRAB	0.721	0.579	0.546	0.514	0.484	0.844	0.671	0.614	0.591	0.536	1.931	13.649

Table 4: Evaluation of conditioned completion and infilling. Lower ADE, FDE, and FID are better; higher Diversity is better.

	ARCTIC		**GRAB**	
Metric	Text2HOI cVAE	MAD-HOI (Ours)	Text2HOI cVAE	MAD-HOI (Ours)
MAE ↓	7.92	12.23	14.75	20.43
MedAE ↓	5.00	8.00	11.20	14.40
Within-4 (%) ↑	44.2	44.3	31.1	33.1
Within-8 (%) ↑	67.2	62.8	43.0	38.3
Interval MAE ↓	0.26	2.32	1.02	4.85
Within Interval (%) ↑	90.7	78.4	80.0	70.5

Table 5: Quantitative evaluation of EOM prediction on ARCTIC and GRAB. We report regression errors, threshold-based accuracies, and distribution-aware tolerance metrics.

Ablation Experiments: We ablate the two design decisions our capability claims rest on: EOM prediction and the learnable modality embeddings that disentangle the three streams. We switch off each individually in our base model and test for atomic generations. As seen in (table 2), modality embeddings are necessary for convergence: without them, FID degrades from 0.071 to 56.7 on ARCTIC and from 1.125 to 11.02 on GRAB, and the network cannot learn a usable distribution. Switching off EOM prediction is comparatively mild but still costly: FID rises from 0.071 to 1.13 on ARCTIC and from 1.125 to 1.47 on GRAB. Geometric metrics also echo similar trends for both datasets. Thus, learning explicit termination does not burden generation and appears to help the model organize varying lengths for similar prompts. The appendix also provides ablations on the geometric losses used for VAE training table 6. We further present results from a generalization experiment with MAD-HOI in Appendix A.7.5.

6 Conclusion

We present MAD-HOI, a masked autoregressive generation framework that produces articulated HOI sequences from text with true handedness. It is further able to generate composite motions, complete and infill motions with minimal degradation in quality. The unique structure of the continuous latent space from our VAE leads to stream-wise modality control and enables learning reliable end-of-motion, which in turn improves generation. In ARCTIC and GRAB, MAD-HOI outperforms retrained open-source baselines on distributional fidelity while being competitive on geometric accuracy and physical plausibility. Finally, its outputs can reliably seed functional retargeting pipelines for humanoid robots (see Appendix A.8 for visualizations).

AI USE STATEMENT

In this work, we used generative AI tools to help develop theoretical models or conceptual frameworks, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results, edit a research paper to improve readability, and identify relevant literature. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, formulate questions for surveys or interviews, create or modify scientific figures or images, suggest experimental parameters and propose a title or keywords for a research paper.

We have reviewed all AI-assisted work. We have cross-checked all text, reported numbers and results, checked LLM-generated research ideas for potential plagiarism through a manual literature survey, references were cross-checked by 1 author, LLM-generated code was verified and tested for correctness by 3 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide all network architecture specifications in table 8 and table 9. Further, all training hyperparameters are detailed in Appendix A.3.4. The main paper and appendix provide details of the modifications to data preparation protocol which is largely borrowed from Cha et al. (2024) and the evaluation protocol from Guo et al. (2022a) We have also submitted a code supplement and will release code publicly upon acceptance.

REFERENCES

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18000–18010, 2023.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, Jiankang Deng, Cordelia Schmid, and Stefanos Zafeiriou. HO-Flow: Generalizable hand-object interaction generation with latent flow matching. *arXiv preprint arXiv:2604.10836*, 2026.

Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36: 30453–30465, 2023.

Jungbin Cho, Junwan Kim, Jisoo Kim, Minseo Kim, Mingu Kang, Sungeun Hong, Tae-Hyun Oh, and Youngjae Yu. Discord: Discrete tokens to continuous motion via rectified flow decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14602–14612, October 2025.

Sammy Christen, Shreyas Hampali, Fadime Sener, Edoardo Remelli, Tomas Hodan, Eric Sauser, Shugao Ma, and Bugra Tekin. Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19888–19901, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 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.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3131–3139, 2025.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pp. 580–597. Springer, 2022b.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910, June 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, Kris Kitani, Matt Feiszli, and Hao Tang. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7136–7146, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. LatentHOI: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17416–17425, 2025a.

Shujia Li, Haiyu Zhang, Xinyuan Chen, Yaohui Wang, and Yutong Ban. Genhoi: Generalizing text-driven 4d human-object interaction synthesis for unseen objects. *arXiv preprint arXiv:2506.15483*, 2025b.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21740–21751, 2024.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 23–40, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72980-5.

Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853*, 2025.

Madhur Mangalam, Sanjay Oruganti, Gavin Buckingham, and Christoph W Borst. Enhancing hand-object interactions in virtual reality for precision manual tasks. *Virtual Reality*, 28(4):166, 2024.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27859–27871, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems (RSS)*, 2023.

Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy˜ human policy. In *Conference on Robot Learning (CoRL)*, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6): 245:1–245:17, 2017.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3352–3360. IEEE, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pp. 346–364. Springer, 2020.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14730–14740, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024a.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27694–27705, 2025a.

Yixuan Zhang, Hui Yang, Chuanchen Luo, Junran Peng, Yuxi Wang, and Zhaoxiang Zhang. Oodhoi: Text-driven 3d whole-body human-object interactions generation beyond training domains. *arXiv preprint arXiv:2411.18660*, 2024b.

Zhenhao Zhang, Ye Shi, Lingxiao Yang, Suting Ni, Qi Ye, and Jingya Wang. Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Zongye Zhang, Bohan Kong, Qingjie Liu, and Yunhong Wang. Towards robust and controllable text-to-motion via masked autoregressive diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9326–9335, 2025c.

Zongye Zhang, Bohan Kong, Qingjie Liu, and Yunhong Wang. Towards robust and controllable text-to-motion via masked autoregressive diffusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9326–9335, 2025d.

A APPENDIX

In this appendix, we provide a review of literature in the larger field of diffusion, jerk profiles for motion reconstructions from discrete and continuous latents, followed by technical details about our method and training (loss definitions, the input token sequence example, training details, and network specifications). We then present details of our evaluation pipeline and metrics, qualitative samples from GRAB, followed by experiments to further validate our claims. Finally, we show our method being applied to a humanoid robot and discuss future work

In the supplementary material, we present interactive 3D visualizations for all our generation capabilities and text to motion from competing baselines. View html file named 'visualize_hoi_motions' under the provided zipped folder. OpenHOI omitted for space constraints and for similar performance to its stem - Text2HOI.

A.1 RELATED WORKS:

Full-body motion generation. Diffusion-based methods such as MDM (Tevet et al., 2023), MotionDiffuse (Zhang et al., 2024a), and MLD (Chen et al., 2023) model continuous action- or text-conditioned trajectories and generate smooth, diverse motions. In parallel, TM2T (Guo et al., 2022b), T2M-GPT (Zhang et al., 2023), MotionGPT (Jiang et al., 2023), and MoMask (Guo et al., 2024) represent motion using discrete VQ tokens, enabling flexible sequence modeling at the cost of quantization. PriorMDM (Shafir et al., 2024) further demonstrates diffusion-based long-duration and spatio-temporal composition, while MARDM (Meng et al., 2025) and MoMADiff (Zhang et al., 2025d) combine continuous diffusion with masked autoregressive generation.

Evolution of diffusion-based sequence modeling. Diffusion models generate continuous data by reversing a progressive noising process (Sohl-Dickstein et al., 2015; Ho et al., 2020). Flow-matching and stochastic-interpolant formulations instead learn a continuous transport field between a simple noise distribution and the data distribution; SiT (Ma et al., 2024), for example, parameterizes this velocity field with a Transformer architecture. Recent work has combined these continuous generative objectives with sequence modeling. Diffusion Forcing (Chen et al., 2024) assigns independent noise levels to causally ordered sequence elements, allowing past tokens to provide cleaner context while future tokens remain more strongly corrupted. MAR (Li et al., 2024), by contrast, extends masked autoregressive modeling to continuous representations: a Transformer predicts contextual features for masked positions, while a diffusion decoder generates the corresponding continuous tokens.

These ideas have subsequently been adapted to motion synthesis. MARDM (Meng et al., 2025) combines bidirectional masked autoregression with a per-token diffusion sampler, while MoMADiff (Zhang et al., 2025c) employs masked autoregressive diffusion for controllable text-to-motion generation. MAD-HOI builds on this continuous masked-autoregressive paradigm, but structures the sequence into separate object, left-hand, and right-hand latent streams and uses a flow-matching head to generate masked motion latents. This formulation enables generation, completion, infilling, modality-specific control, and learned termination within a single model.

A.2 TOKENIZATION: DISCRETE VS CONTINUOUS

We argue that motions reconstructed from discrete codes lose fine-grained information and this directly results in motions that are less smooth and more jittery. We take inspiration from (Cho et al., 2025), and assess the temporal coherence of the reconstructed motions by analyzing their jerk profiles, i.e., the third derivative of position over time. Jerk quantifies fine-grained temporal irregularities, allowing us to distinguish overly static motion from high-frequency jitter. Lower and smoother jerk magnitudes correspond to more natural and physically consistent trajectories. See fig. 4 and fig. 5 for plots on jerk profiles of ground truth and reconstructed trajectories of the wrist joints and the object centers obtained from the VAE and its equivalent VQ-VAE for samples from ARCTIC and GRAB respectively. We observe that the VQ-VAE-reconstructed trajectories are less smooth, with spikes corresponding to abrupt, non-physical changes. The trajectories from the VAE are closer in smoothness to the ground truth data.

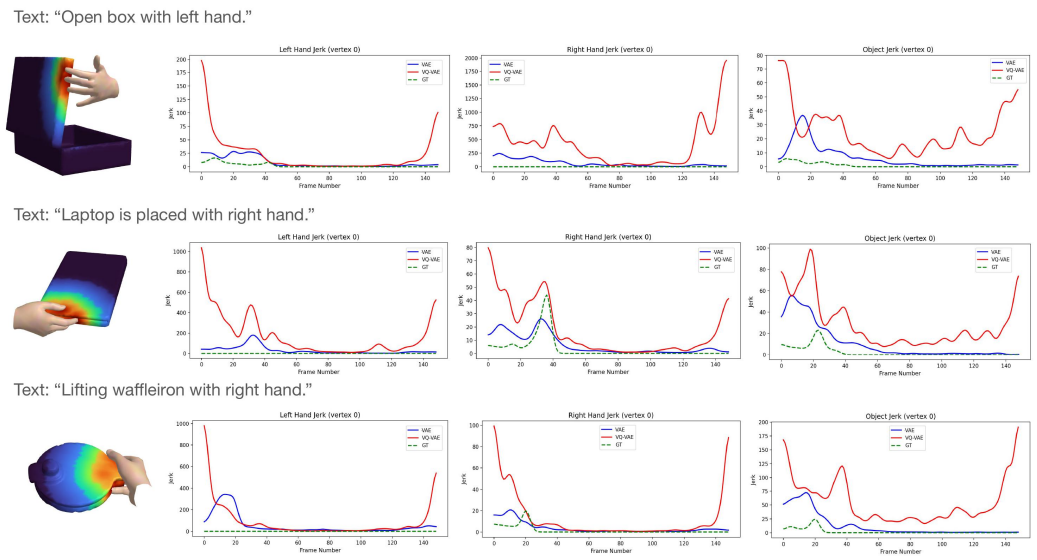


Figure 4: VAE vs VQ-VAE Reconstruction Jitter for ARCTIC: Jerk Analysis for Smoothness Comparison. We plot jerk magnitude for ground truth, VAE-reconstructed, and VQ-VAE-reconstructed trajectories of the MANO hand wrist joints and object centers.

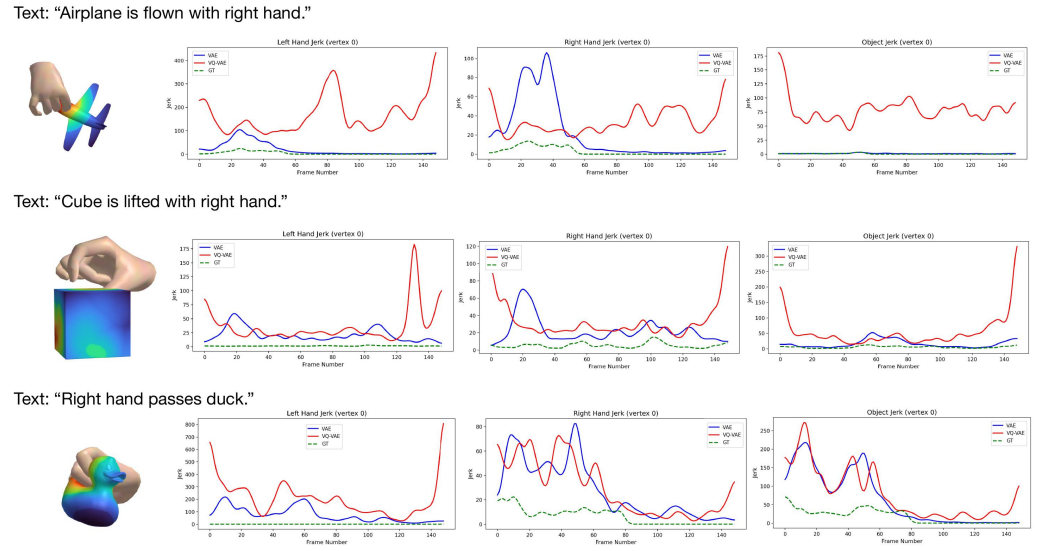


Figure 5: VAE vs VQ-VAE Reconstruction Jitter for GRAB: Jerk Analysis for Smoothness Comparison. We plot jerk magnitude for ground truth, VAE-reconstructed, and VQ-VAE-reconstructed trajectories of the MANO hand wrist joints and object centers.

Metric	w/o $\mathcal{L}_{\text{Pen}}$	w/o $\mathcal{L}_{\text{C}}$	w/o $\mathcal{L}_{\text{DM}}$	**All losses**
Obj. L1 ↓	0.005	0.010	0.018	**0.005**
Hand L1 ↓	0.013	0.023	0.076	**0.013**
MPJPE ↓	0.049	0.052	0.119	**0.044**
Pen. (%) ↓	61.16	34.94	0.00	**2.10**
Con. (%) ↑	98.83	79.48	22.4	**92.20**

Table 6: **VAE Training on GRAB**: Ablation of the geometric loss terms .

A.3 Method Details

A.3.1 Losses for VAE Training

Our VAE is trained to reconstruct fine-grained motions and hand-object contacts. To facilitate this, we apply some geometric losses that encourage close contact and penalize penetration. They are defined as follows:

Contact Loss: We enforce a threshold-based contact loss. $D(J_i^o)$ calculates the distance between a hand joint and the closest object surface, enforcing the threshold distance constraint $D(J_i^o) \leq \phi$:

$$\mathcal{L}_C = \sum_i D(\hat{J}_i^o), \quad \forall\, D(\hat{J}_i^o) \leq \phi, \tag{4}$$

where

$$D(\hat{J}_i^o) = \min_j \left\| V_j^p - \hat{J}_i^o \right\|_2^2, \tag{5}$$

with V_j^p representing object surface points and J_i^o representing hand joints.

Penetration Loss: To penalize penetration of the hand mesh and the object mesh, we minimize the average squared distances of penetrating hand vertices $p \in P_{\text{in}}^o$ to their closest object vertex $\hat{V}_i{}^o$:

$$\mathcal{L}_{\text{pen}} = \frac{1}{|P_{\text{in}}^o|} \sum_{p \in P_{\text{in}}^o} \min_i \|p - \hat{V}_i{}^o\|_2^2. \tag{6}$$

Distance Map Loss: Following (Liang et al., 2024), we apply a distance map loss to align the estimated distance map (distance from each hand joint to the closest object vertex) with the ground-truth distance map. For a single hand,

$$\mathcal{L}_{\text{DM}} = \sum_i \big(D(\hat{J}_i^o) - D(J_i^o)\big)^2 \, \mathbf{1}\big[D(J_i^o) < \phi\big], \tag{7}$$

where ϕ is a threshold that selects near-contact regions.

We also study the effect of these losses on the reconstruction quality of our VAE. See table 6 to see the ablated performance of our VAE on the GRAB dataset. Without the penetration loss, there is high contact but it it assuredly comes with penetration. Without the contact loss, the contact % is lower than optimum. Perhaps the most important loss is the Distance Map Loss, as when its removed the contact worsens severely, and further the reconstruction losses and MPJPE is also impacted. The combination of all losses performs the best.

A.3.2 Input Token Sequence to Transformer Example

If we assume a max token length of 18 (6 possible latent timesteps x 3 for modalities), and consider a motion having 5 valid latent timesteps, where we mask one latent timestep out and the EOM tokens are appended at the end, the sample input token sequence to the transformer would look like eq. (8).

$$\langle z_o^0 \rangle \langle z_l^0 \rangle \langle z_r^0 \rangle \langle z_o^1 \rangle \langle z_l^1 \rangle \langle z_r^1 \rangle \langle \text{MASK} \rangle \langle \text{MASK} \rangle \langle \text{MASK} \rangle \langle z_o^3 \rangle \langle z_l^3 \rangle \langle z_r^3 \rangle \langle z_o^4 \rangle \langle z_l^4 \rangle \langle z_r^4 \rangle \langle \text{EOM}_o \rangle \langle \text{EOM}_l \rangle \langle \text{EOM}_r \rangle \tag{8}$$

Layer	Operation	Output
Input	Object sequence	$B \times T \times C_{obj}$
Preprocess	permute$(B, T, C_{obj}) \to (B, C_{obj}, T)$	$B \times C_{obj} \times T$
Encoder	Conv1d$(C_{obj} \to D, k{=}3, s{=}1, p{=}1)$ + ReLU	$B \times D \times T$
Encoder block	Conv1d$(D \to D, k{=}4, s{=}2, p{=}1)$ + ResNet1D	$B \times D \times T/2$
Encoder block	Conv1d$(D \to D, k{=}4, s{=}2, p{=}1)$ + ResNet1D	$B \times D \times T_{lat}$
Latent proj.	Conv1d$(D \to D, k{=}3, s{=}1, p{=}1)$	$B \times D \times T_{lat}$
Posterior mean	Conv1d$(D \to D, k{=}1)$	$\mu:\ B \times D \times T_{lat}$
Posterior log-var	Conv1d$(D \to D, k{=}1)$, clamp $[-10, 10]$	$\log \sigma^2:\ B \times D \times T_{lat}$
Sampling	$z = \mu + \epsilon \odot \exp(\frac{1}{2} \log \sigma^2)$	$B \times D \times T_{lat}$
Concat	concat$(z, f_{pointnet})$	$B \times (D{+}P) \times T_{lat}$
Decoder	Conv1d$(D{+}P \to D)$ + ReLU	$B \times D \times T_{lat}$
Decoder block	ResNet1D $\to$ Upsample $\to$ Conv1d	$B \times D \times T/2$
Decoder block	ResNet1D $\to$ Upsample $\to$ Conv1d	$B \times D \times T$
Decoder	Conv1d$(D \to D)$ + ReLU	$B \times D \times T$
Output	Conv1d$(D \to C_{obj})$	$B \times C_{obj} \times T$
Postprocess	permute$(B, C_{obj}, T) \to (B, T, C_{obj})$	$B \times T \times C_{obj}$

Table 7: Network architecture of the object stream of the VAE.

A.3.3 Flow-Matching Details

Given the linear interpolant $x_t = \alpha_t x_0 + \sigma_t \epsilon$ from the main paper, with $\alpha_t = 1 - t$ and $\sigma_t = t$, the velocity field toward which the diffusion MLP regresses is defined as

$$v(x, t) = \dot{\alpha}_t \, \mathbb{E}[x_0 \mid x_t = x] + \dot{\sigma}_t \, \mathbb{E}[\epsilon \mid x_t = x], \tag{9}$$

where $\dot{\alpha}_t$ and $\dot{\sigma}_t$ are the derivatives of the interpolation coefficients with respect to t.

A.3.4 Training Details

We train the generation model on 4 H100 GPUs for 200000 iterations for ARCTIC and 150000 iterations for GRAB with a max learning rate of 3e-4, which is achieved through a linear warmup over 1000 iterations. The learning rate then decays with cosine annealing to 1e-4. The VAE is trained on 2 H100 GPUs for 6000 epochs with a learning rate of 2e-4, with $\alpha = 1$, $\beta = 0.5$, $\gamma = 0.5$, $\delta = 1$. The KL term is internally scaled by $\beta_{\mathrm{KL}} = 10^{-4}$ and subsequently weighted by $\lambda_{\mathrm{KL}} = 0.5$. We use this small coefficient because the KL divergence is summed over the latent-channel dimension before being averaged over valid temporal tokens. Our VAE operates over short temporal sequences, as opposed to static poses. We use a window length of 152 to maximize the temporal context within the VAE and have a 4 frames to 1 token (latent) ratio throughout the method.

The longer window length presents some challenges when dealing with a dataset like ARCTIC, which has many shorter atomic motion sequences. For shorter sequences, we pad the last valid frame until the end and also maintain replicate padding within the VAE to ensure the network does not see zeros or abrupt signal changes, which could lead to drift-like artifacts during decoding. Further, the VAE reconstruction losses are computed over the padded sequences as opposed to the valid-only frames, forcing the VAE to learn to reconstruct static hands and objects from idle latents. This design complements the EOM mechanism (see main paper section 4.2).

A.4 Network Architecture Specifications

A.4.1 VAE

Shape symbols: B: batch size, T: input number of frames, T_{lat}: number of tokens where $T_{lat} = T/2^{down_t}$ and `down_t`=2, C_obj: channels in object feature (10 for ARCTIC and 9 for GRAB), C_hand: channels in hand feature (100), D: latent dimension (512), P: channels in pointnet features (1024), mu: distribution mean and σ: log variance. See table 7 and table 8 for the specification tables.

Layer	Operation	Output
Input	Left / right hand sequences	$B \times T \times C_{hand}$
Conditioning	Object latent z_{obj}, object reconstruction $\hat{x}_{obj}$, PointNet features $f_{pointnet}$	—
Preprocess	permute$(B, T, C_{hand}) \to (B, C_{hand}, T)$	$B \times C_{hand} \times T$
Shared encoder	Conv1d$(C_{hand} \to D, k{=}3, s{=}1, p{=}1)$ + ReLU	$B \times D \times T$
Encoder block	Conv1d$(D \to D, k{=}4, s{=}2, p{=}1)$ + ResNet1D	$B \times D \times T/2$
Encoder block	Conv1d$(D \to D, k{=}4, s{=}2, p{=}1)$ + ResNet1D	$B \times D \times T_{lat}$
Latent proj.	Conv1d$(D \to D, k{=}3, s{=}1, p{=}1)$	$B \times D \times T_{lat}$
Distribution mean	Conv1d$(D \to D, k{=}1)$	$B \times D \times T_{lat}$
Distribution log-variance	Conv1d$(D \to D, k{=}1)$, clamp $[-10, 10]$	$B \times D \times T_{lat}$
Sampling	reparameterized sample $\to z_{lhand}, z_{rhand}$	$B \times D \times T_{lat}$ each
Concat	concat$(z_{obj}, z_{lhand}, z_{rhand}, f_{pointnet})$	$B \times (3D{+}P) \times T_{lat}$
FiLM cond.	AvgPool1d$(k{=}s,$ stride${=}s)$ on $\hat{x}_{obj} \to$ Conv1d$(C_{obj} \to 2(3D{+}P), k{=}1)$	$B \times 2(3D{+}P) \times T_{lat}$
FiLM split	split into (γ, β)	$2 \times [B \times (3D{+}P) \times T_{lat}]$
FiLM apply	$\gamma \odot z + \beta$	$B \times (3D{+}P) \times T_{lat}$
Decoder	Conv1d$(3D{+}P \to D)$ + ReLU	$B \times D \times T_{lat}$
Decoder block	ResNet1D $\to$ Upsample $\to$ Conv1d	$B \times D \times T/2$
Decoder block	ResNet1D $\to$ Upsample $\to$ Conv1d	$B \times D \times T$
Decoder	Conv1d$(D \to D)$ + ReLU	$B \times D \times T$
Output	Conv1d$(D \to 2C_{hand})$	$B \times 2C_{hand} \times T$
Postprocess	permute$(B, 2C_{hand}, T) \to (B, T, 2C_{hand})$	$B \times T \times 2C_{hand}$

Table 8: Network Architecture of the hand stream of the VAE.

A.4.2 GENERATOR

For our generator, we use a single attention block with 16 heads and follow it up with an SiT-XL network which has 16 residual blocks. The flow matching head is queried for 18 refinement steps where random tokens are denoised repeatedly.

Shape symbols: B: batch size, L: sequence length (number of max motion tokens = 117; 38 tokens x 3 modalities + 3 EOM tokens), D_{ae}: AE latent dimension (512), D: transformer hidden dimension (1024), D_{ff}: transformer feed-forward hidden dimension (4096), H: number of attention heads (16), N_{tr}: number of MAR transformer blocks (1), N: number of masked/selected tokens processed by DiffMLPs (variable), W_{diff}: DiffMLPs hidden width (1792), N_{diff}: number of DiffMLPs residual blocks (16). See table 9 for the specification table.

A.5 EVALUATION PIPELINE

We evaluate MAD-HOI with a learned text-motion evaluator adopted from HumanML3D (Guo et al., 2022a), following the standard practice of measuring generated motions in a semantic embedding space rather than directly in the raw motion parameter space. We train an evaluator with real hand-object interaction data to learn aligned representations for text and motion.

The evaluator consists of a motion representation module, a text encoder, and a motion encoder. On the motion side, we first compress the full hand-object motion sequence using a temporal convolutional movement encoder. This module reduces the temporal resolution and produces a compact latent motion representation that captures the essential dynamics of the sequence. A corresponding decoder is used only during pre-training of this motion representation. On the text side, we encode tokenized captions using pre-trained Word2Vec embeddings together with POS-style linguistic indicators, which are passed through a bidirectional GRU and projected into a shared embedding space. On the motion side, the compact motion latents are processed by a second bidirectional GRU and projected into the same embedding space. In the first stage of training, we learn the motion representation using an auto-encoding objective on real hand-object motion sequences. The model reconstructs the input sequence from its latent representation, while regularization terms encourage sparse and temporally smooth motion latents. This stage provides a compact and stable motion representation that is later used for text-motion alignment. In the second stage of training,

Layer	Operation	Output Features
A. Top-Level MARDM Pipeline		
Input	Latent token sequence	$B \times L \times D_{ae}$
InputProcess	$\mathrm{Linear}(D_{ae} \to D)$	$B \times L \times D$
Time + Modal Embedding	Add time and modality embeddings	$B \times L \times D$
Transformer Stack	$\mathrm{MARTransBlock} \times N_{tr}$	$B \times L \times D$
Diffusion Head	DiffMLPs_SiT on masked tokens	$B \times L \times D_{ae}$
B. MARTransBlock (Detailed)		
AdaLN Modulation	$\mathrm{SiLU} \to \mathrm{Linear}(D \to 6D)$, split to shifts/scales/gates	$6 \times (B \times D)$
Multi-Head Self-Attention (MSA)	Q/K/V projections, scaled dot-product attention, output projection	$B \times L \times D$
MLP	$\mathrm{Linear}(D \to D_{ff} \to D)$ with GELU	$B \times L \times D$
Residual Gating	Gated residual updates for MSA and MLP branches	$B \times L \times D$
C. DiffMLPs_SiT (Detailed)		
Input/Cond Proj	$\mathrm{Linear}(D_{ae} \to W_{diff})$, $\mathrm{Linear}(D \to W_{diff})$	$N \times W_{diff}$
Time Embedder	Sinusoidal timestep embedding + MLP	$N \times W_{diff}$
Res Stack	$\mathrm{ResBlock} \times N_{diff}$ with adaLN gating	$N \times W_{diff}$
Final Layer	AdaLN-modulated LN + $\mathrm{Linear}(W_{diff} \to D_{ae})$	$N \times D_{ae}$
Sampling	ODE sampler	$N \times D_{ae}$

Table 9: Unified MAD-HOI generator architecture.

we freeze the motion representation module and train the text encoder and motion encoder jointly with a contrastive objective. For each training sample, the caption and its paired motion are treated as a positive pair and are encouraged to have similar embeddings. Negative pairs are formed by mismatching captions and motions within the batch, so the evaluator learns to distinguish semantically correct interactions from incorrect ones.

For text to motion evaluation, we generate motions from text prompts from MAD-HOI as well as baselines, and feed them to the frozen evaluator to compute all generation metrics. Both generated and real motions are mapped in the evaluators' motion embedding space, while captions are mapped in the corresponding text embedding space. We compute retrieval-style and distribution-level metrics in this shared space.

A.6 Metrics

A.6.1 Generation/Distribution-based Metric Definitions

Top-3 Accuracy (R@3) measures how often the correct conditioning text is ranked among the three closest matches when retrieving in the shared embedding space between motions and conditions. Higher R@3 indicates higher text–motion semantic alignment. Frechet Inception Distance (FID) measures the discrepancy between the distributions of real and generated motion embeddings by comparing their means and covariance matrices under a Gaussian approximation. Kernel Inception Distance (KID) measures the discrepancy between real and generated motion embeddings using a kernel-based maximum mean discrepancy estimator. Unlike FID, it does not assume Gaussian feature distributions and is less biased for small sample sizes. Lower FID and KID indicate greater similarity between the generated and real motion distributions. Diversity is the variance within features extracted from the generated motions, and Matching Score is the embedding distance between generated motions and their conditioning text. Diversity values closer to the ground-truth diversity indicate more realistic coverage of the data distribution and lower Matching Scores indicate stronger text–motion semantic alignment. We omit Multimodality because its pairwise-distance formulation can reward unrestricted or implausible variation and lacks a corresponding ground-truth reference; diversity is instead assessed alongside distributional, semantic-alignment, physical-plausibility, and human-evaluation metrics. All our generation metrics are averaged over 5 evaluation repeats with varying seeds. The standard deviations were negligible and did not affect the ranking of methods; therefore, they are omitted for readability. The geometric metrics computed once only.

A.6.2 COMPLETION AND INFILLING METRICS:

Average Displacement Error (ADE) measures how closely the best generated motion matches the ground-truth motion over the entire sequence by averaging the displacement error across all time steps. Final Displacement Error (FDE) measures the displacement between the final pose of the ground-truth motion and the final pose of the closest generated sample. Lower ADE and FDE indicate better agreement with the ground-truth trajectory and endpoint, respectively, and are common metrics for evaluating sample accuracy in trajectory forecasting (Yuan & Kitani, 2020).

A.6.3 EOM METRICS:

We evaluate End-of-Motion (EOM) prediction using both standard regression metrics and distribution-aware tolerance metrics. The mean absolute error (MAE) measures the average absolute difference between the predicted and ground-truth sequence length (number of frames), while the median absolute error (MedAE) captures the typical error and is less sensitive to outliers. We also report Within-4 and Within-8 $\%$s, which denote the fraction of predictions that fall within tolerance thresholds of 4 and 8 frames from the ground truth sequence lengths, respectively. Since a single text prompt can correspond to multiple valid sequence lengths (from different human demonstrations) in the dataset, we additionally use distribution-aware metrics based on an interval $[q_{10}, q_{90}]$ for each unique condition (which is a [action, object, left_use, right_use] tuple). Interval MAE measures the distance to this valid range, assigning zero penalty when the prediction lies inside it, while Within Interval reports the proportion of predictions that fall within the semantically plausible duration range from the ground truth sequences.

In reference to table 5, MAD-HOI achieves strong and semantically meaningful EOM performance across both datasets, though it does not beat the Text2HOI cVAE. On ARCTIC, MAD-HOI attains an MAE of 12.23 ± 0.88 frames and a Median AE of 8.00, showing that its predicted stopping point is typically close to the true end of the motion. On GRAB where the problem is more challenging overall, MAD-HOI attains a higher MAE of 20.43 ± 1.09 frames and a Median AE of 14.4. Importantly, the distribution-aware metrics are particularly encouraging: an Interval MAE of 2.32 ± 0.50 for ARCTIC and 4.85 ± 0.75 for GRAB show that when we deviate from the distribution, we observe errors within 2-5 frames. The Within Interval %s of $78.4\% \pm 1.3\%$ for ARCTIC $70.5\% \pm 1.3\%$ for GRAB and show that most MAD-HOI predictions fall inside the valid duration range for the corresponding semantic interaction, even when there is substantial natural variability in demonstration length.

A.7 EXPERIMENTS

We visualize sample motions from ARCTIC and GRAB datasets with time-progressing grasps for MAD-HOI and the baseline methods in fig. 6 and fig. 7. We observe that generations from our method align semantically better with the text prompts than baselines, both in terms of motion and hand use. We also observe that our method minimizes penetrations while maintaining consistent contact. Please see supplementary material for interactive motion samples from our method and baselines.

A.7.1 USER PREFERENCE

We conduct a small user-study to evaluate human preferences for our method and the baselines. We evaluate the five generation methods (MAD-HOI, DiffH2O, Text2HOI, OpenHOI, LatentHOI) through a forced-ranking study on ARCTIC (articulated objects) and GRAB (rigid objects). Each trial presents one text prompt alongside five anonymised, interactive 3D viewers labelled A–E, with the slot-to-method assignment randomised independently per trial so that no letter is associated with a method. Participants produce two complete 1–5 rankings per trial: motion–text alignment — how faithfully the motion depicts the described action — and physical plausibility — naturalness of hand–object contact, penetration and object motion, judged independently of the text. All methods are rendered through an identical pipeline, with the same camera treatment, the same colour coding for left and right hands, and the same playback rate, so that no method is visually distinguishable from another. Each participant completes 50 trials, with prompts drawn without replacement from ARCTIC and GRAB prompts, and we collect three raters per dataset.

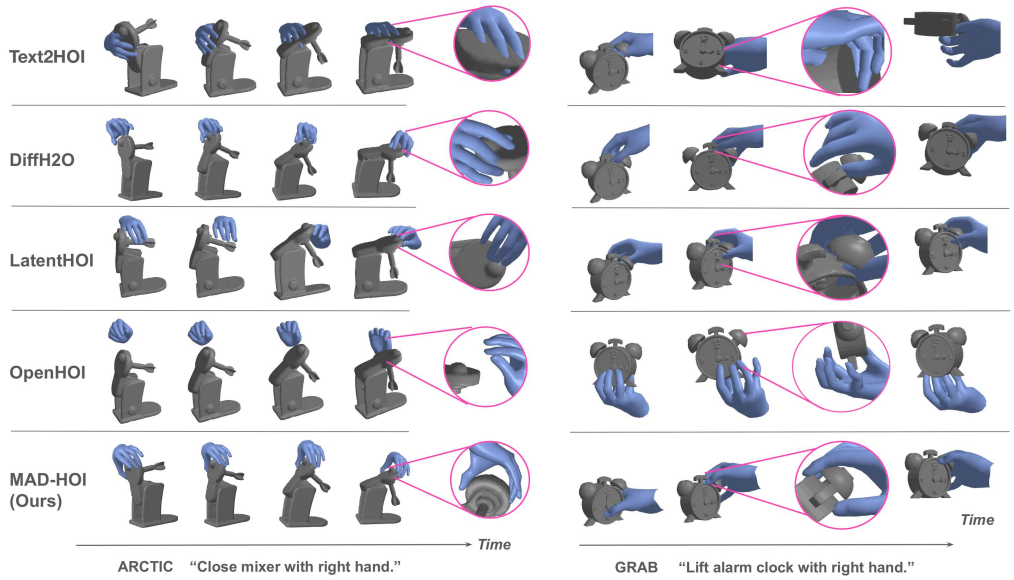


Figure 6: **Comparison of time-progressing grasps with baselines**. One frame is visualized with zoom-in to highlight contact and penetrations in generations across all methods. Right hand visualized in blue.

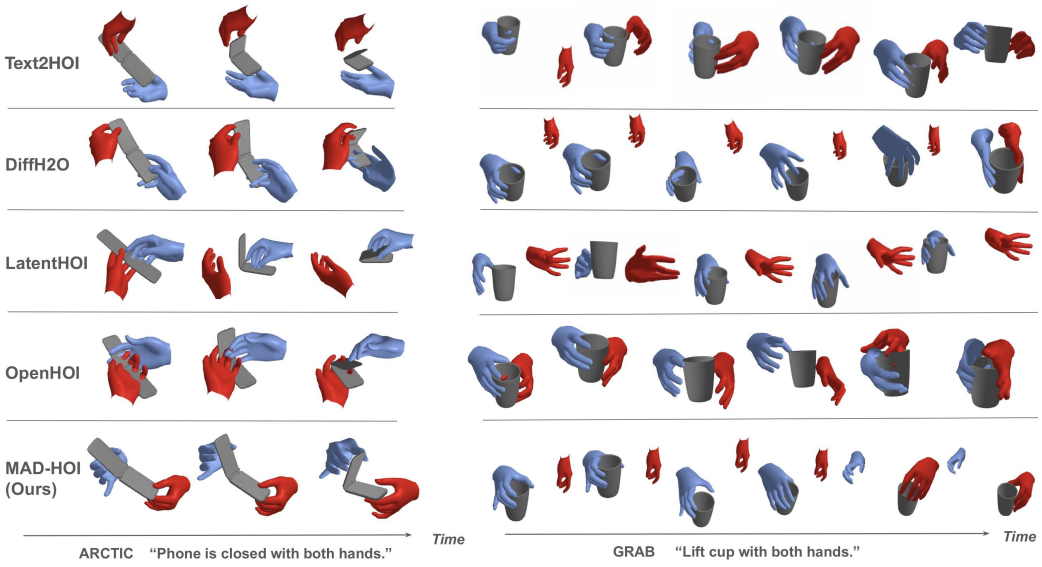
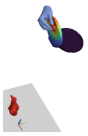


Figure 7: **Comparison of time-progressing grasps with baselines**. Right hand visualized in blue; Left hand visualized in red.

MAD-HOI attains the best mean rank on both criteria — 2.26 for motion–text alignment and 2.26 for physical plausibility (1 = best of five, 3.0 = chance) — ahead of DiffH2O (2.48 / 2.50), Text2HOI (2.81 / 2.91), OpenHOI (3.51 / 3.58) and LatentHOI (3.95 / 3.75), and is ranked first in 43.3% of trials on both axes. Separation from Text2HOI, OpenHOI and LatentHOI is substantial, spanning 0.55 to 1.69 rank positions. DiffH2O is the closest competitor: in direct comparison MAD-HOI is preferred in 54.0trials for alignment and 55.3% for plausibility. The study therefore indicates that MAD-HOI is preferred over three of the four baselines on both criteria, and is competitive with DiffH2O.

A.7.2 VALIDATING HANDEDNESS: QUANTIFYING 'UNUSED HAND' MOTION IN MAD-HOI

In order to quantify the claim of our method learning to 'switch-off' unused hands, we present table A.7.2. It evaluates the 'generated motion' for inactive hands in prompts specifying a single hand. The wrist and vertex displacement are in m, compounded over the entire sequence. The hand accuracy calculates the % of frames where the unsued hand wrist (defined by prompt) moves less than 1 cm. Our idle latents corresponding to 'unused hands' are forced to predict zeros. And since neural networks learn smooth approximations, obtaining zero-latents for unused hands inevitably has some noise. This noise does lead to some non-zero cumulative joint motion over all the frames of a sequence, and therefore, the displacement errors are not zero. However, the smaller inactive hand displacement MAD-HOI has compared to Text2HOI confirms that our method predicts near-zero motion for unused hands, and the presence of explicit latents for unused hands in our VAE. We also present a visual depiction of a frame from the generated motion for 'Close the waffleiron with right hand' from MAD-HOI. Here, the unused hand - the left hand - shown without any masking is truly inactive, while the right hand completes the desired motion. The generated poses for this hand are near-zeros, leading to the twisted MANO hand mesh.



	ARCTIC		GRAB	
Idle metric	**T2H**	**Ours**	**T2H**	**Ours**
Wrist disp. ↓	0.068	**0.030**	0.274	**0.003**
Vertex disp. ↓	0.078	**0.033**	0.331	**0.047**
Handedness Acc. ↑	0.54	**0.64**	0.36	**1.00**

Table 10: Idle-hand comparison between Text2HOI and MAD-HOI. Inactive-hand wrist and vertex displacement and handedness accuracy on ARCTIC and GRAB.

Learning to independently model single-handed and bimanual motions without post-hoc masking is important when the interaction looks very different when being carried out with one hand vs both hands. For example, the grasp that a hand might make while opening a box might vary depending on whether a second hand stabilizes the box. Learning to differentiate single-handed motions from bimanual ones can create modes in the latent space which could be the difference between stable and plausible motions that actually achieve the desired object articulation. This is especially important when such text-to-motion models are applied to embodied use cases like robotics.

A.7.3 LEARNED EOM VS USER-INPUT LENGTH

We observed that predicting the EOM token helps MAD-HOI generate better motions. However, we argue that allowing the method to jointly estimate sequence length is better than using user-specified length. Our method, like our baselines, is not trained to generate motions at varying frame rates, and therefore, if a user input for sequence length is very different from lengths the model has seen during training, it could harm the quality of the generated motion. The assumption that the user has greater controllability by specifying sequence length hinges on the user being an expert and knowing the details of the training data. To demonstrate how 'wrong' length inputs can harm generation quality, we conduct an experiment where we force the learned EOM token to be at -4 frames, -8 frames and -12 frames from the ground truth lengths and compare the generated motions against the model-predicted EOM baseline. Forcing lengths greater than GT is not effective with our MAD-HOI baseline here as the method will predict an EOM token per its learned distribution and this will likely come before the user-specified EOM token. As seen in table 10, we observe that the farther the user-specified length is from the ground truth length, the worse the quality of generated motion. The baseline performance - detected EOM in MAD-HOI - reflects the learned natural distribution of motions. When allowed to detect the sequence end, the model is able to generate more natural and expressive motions that are semantically aligned with the textual intent.

A.7.4 BASELINES FOR COMPLETION AND INFILLING

We compare MAD-HOI against simple, protocol-matched, non-learned kinematic baselines. For completion, we extrapolate the motion at the constant velocity given by the last two observed frames, extrapolating rotations on $\mathrm{SO}(3)$. For infilling, we interpolate between observed frames, using linear interpolation for translational and articulation components and SLERP (Spherical Linear Interpolation) for rotations; gaps observed on only one side hold the nearest observed frame, and sequences

Length	FID ↓	KID ↓	Div.	Top-1 ↑	Top-3 ↑	Match ↓	Pen. %	Contact %
−12 frames	4.240	0.0595	19.087	0.699	0.882	4.616	17.13	79.89
−8 frames	0.623	0.0119	19.997	0.817	0.975	2.066	17.78	81.37
−4 frames	0.130	0.0035	20.037	0.837	0.991	1.273	18.08	81.58
MAD-HOI detected end	0.075	0.0004	20.130	0.837	0.993	1.097	15.25	83.86

Table 10: Forced EOM sequence-length ablation on ARCTIC. Motions are generated with the supplied length shortened by 4, 8, and 12 frames relative to the MAD-HOI detected end-of-motion.

with no observed frame within their valid length fall back to the per-object mean pose computed from training sequences. For ADE/FDE, all methods are evaluated on the same test sequences, with the same masks and the same masked frames (for MAD-HOI, those within its predicted sequence length).

The kinematic baselines provide a strong test of whether MAD-HOI learns interaction dynamics beyond simple local continuation. From section A.7.4, we see that for completion, MAD-HOI substantially outperforms constant-velocity extrapolation on GRAB across all ADE and FDE settings. On the shorter ARCTIC sequences, constant-velocity extrapolation becomes competitive as more context is provided, reflecting the effectiveness of a local motion prior over short prediction horizons; nevertheless, MAD-HOI achieves fetter ADE and FDE when lesser context is provided and more completion is necessary (5 context tokens provided does shorten the number of frames left in the motion that need to be completed). The FID and Diversity scores also underline how MAD-HOI outperforms the constant velocity extrapolation.

For infilling, MAD-HOI consistently outperforms interpolation in ADE and FDE on both ARCTIC and GRAB, while also achieving substantially lower FID. These results show that the masked generative model does more than interpolate between observed poses: it learns interaction dynamics that better reconstruct the missing motion while remaining close to the distribution of valid HOI trajectories

			ADE ↓					FDE ↓					FID ↓	Diversity ↑
Task Tokens (#):			1	2	3	4	5	1	2	3	4	5	2	2
Completion	ARCTIC	Const. Vel.	0.733	0.563	0.428	0.304	0.272	1.532	1.209	0.933	0.656	0.589	3.873	18.823
		MAD-HOI	0.649	0.617	0.613	0.608	0.592	0.816	0.787	0.789	0.737	0.733	0.173	20.325
	GRAB	Const. Vel.	2.911	2.909	3.231	3.423	3.473	4.928	5.197	6.415	7.260	7.612	14.546	11.174
		MAD-HOI	0.748	0.776	0.738	0.749	0.724	0.963	1.019	1.001	0.960	0.882	1.719	13.010
Infilling	ARCTIC	Interp.	1.284	1.135	1.026	0.953	0.869	1.325	1.158	1.041	0.951	0.856	7.025	18.582
		MAD-HOI	0.783	0.707	0.675	0.639	0.629	0.833	0.766	0.719	0.691	0.684	0.182	19.214
	GRAB	Interp.	1.224	0.917	0.764	0.651	0.545	1.337	1.209	1.077	0.942	0.842	4.582	12.707
		MAD-HOI	0.721	0.579	0.546	0.514	0.484	0.844	0.671	0.614	0.591	0.536	1.931	13.649

Table 11: Conditioned completion and infilling against non-learned kinematic baselines. ADE/FDE are computed on the predicted frames over hand-joint trajectories (21 joints per active hand) and the object position trajectory, each z-scored with dataset statistics and averaged with equal weight (unitless). Const. Vel.: constant-velocity extrapolation from the last two observed frames (rotations on SO(3)) till GT sequence length. Interp.: linear/SLERP interpolation between observed frames. Lower ADE, FDE, and FID are better; higher Diversity is better.

A.7.5 GENERALIZATION

In order to test the robustness of MAD-HOI, we design another split for the ARCTIC dataset where we exclude certain motion samples with single handedness (ex. *'Close Laptop with left hand', 'Right hand opens mixer'* from the training set and evaluate the trained model on these excluded samples to test generalization of motion transfer between hands. Our base ARCTIC split is termed ***'Level 1'*** and the split with these exclusions is termed ***'Level 2'***. We provide metrics for text to motion on both splits in table 12. As the performance of our model on the Level 2 ARCTIC split is not much worse than Level 1, we conclude that our modality tokens help the model generalize to out-of-distribution motions regarding *'handedness'*.

Split	Accuracy (Top 3) ↑	FID ↓	KID ↓	Diversity →	Matching ↓	Accel. (Jitter) ↓	Penetration (%) ↓	Contact (%) ↑
Ground Truth	0.991	-	-	20.235	1.295	0.187	7.864	80.771
Level 1	0.944	0.071	0.004	20.312	1.353	0.18	12.89	84.47
Level 2	0.914	0.137	0.006	20.717	1.525	0.299	13.012	80.637

Table 12: Quantitative evaluation of text to motion for Level 1 and Level 2 splits from ARCTIC. Ground truth values are provided for reference and are highlighted in gray.

When we tested our method for similar splits which exclude certain 'actions' or 'objects', we did not see robust generalization and this is expected as the *'action'* signal is primarily encoded by CLIP embeddings and the *'object geometry'* is encoded by pointnet features only. Therefore, for more intent-based or geometry-based generalization, we require richer descriptors and perhaps training on much larger datasets which can offer a diverse grasp prior.

A.8 Application

High-fidelity 4D HOI sequences can serve as priors for robotic motion by capturing detailed human-object interactions. These sequences provide smooth trajectories with realistic contact events and limited penetration between bodies, properties that matter when learning control policies. Furthermore, text-driven generation of these sequences reduces the dependence on manually collected human demonstrations, which otherwise require extensive and labor-intensive teleoperation setups (Khazatsky et al., 2024; O'Neill et al., 2024; Qiu et al., 2025).



Figure 8: **Robot retargeting from MAD-HOI outputs.** A text-generated HOI sequence is retargeted from the MANO model to the Unitree G1 robot with Inspire dexterous hands. The resulting trajectory preserves realistic contact and grasp transitions, indicating that our generated motions can serve as reference priors for robotic manipulation.

Although an embodiment gap exists between the MANO model and dexterous robot hands, we find that our generated sequences can seed a manipulation pipeline. We bridge this gap in multiple stages: first, the MANO motion is retargeted to the robot hand using keypose-vector optimization (Qin et al., 2023). This target motion is then translated to the full robot using Inverse Kinematics. Finally, this reference trajectory is refined using the physics-based rewards of DexMachina's (Mandi et al., 2025) functional retargeting pipeline to produce a fine-tuned policy. Fig. 8 shows a robot trajectory, with motion retargeted from our generated sequence to the Unitree G1 robot with Inspire hands.

A.9 Future Work

While we verify generalization across motion handedness, generalizing to unseen objects and actions will require pretraining on large datasets. Whether this pretraining would be more effective with static grasps (much higher scale of data available but lacks temporal context) or short motion priors (temporal data but scarce), is unknown and can be studied in more detail. Perhaps a two-stage generation setup with the first stage focusing on static grasps and the second stage focusing on

motion could work well. However, as seen in our baselines that adopt this type of architecture, such a design is better suited for motions and objects that do not require articulations (DiffH2O and LatentHOI work better on GRAB than ARCTIC).