CAD-WAM: DEXTEROUS MANIPULATION GENERATION VIA CONTACT-AWARE WORLD ACTION MODEL

Anonymous authors
Paper under double-blind review

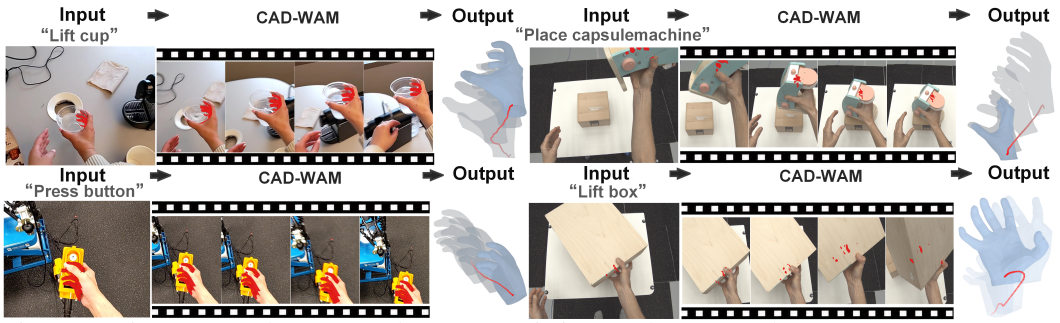


Figure 1: Given a textual prompt and an egocentric image, the proposed CAD-WAM generates realistic and vivid 4D hand interaction trajectories. *Left*: the input of CAD-WAM. *Middle*: **Imagined key *video* frames and *contacts* rollout** *Right*: The generated 4D hand interaction trajectories at $t = 8, 16, 24, 32$ steps. The red gradient curve denotes the wrist-joint trajectory over time.

ABSTRACT

Generating 4D hand interaction trajectories from a single egocentric image and a language instruction requires anticipating how hand–object contact evolves across different hand regions. Most existing methods map vision and language directly to actions without explicitly imagining how the interaction unfolds, and their global conditions leave fine-grained contact under-constrained, so contact emerges only as an implicit by-product of global motion. We present **CAD-WAM**, a contact-aware world action model for dexterous generation which jointly imagines future egocentric video and hand actions, grounding trajectory generation in imagined interaction dynamics rather than static observations. Our key idea is to recast regional contact evolution from an implicit consequence of global motion into an explicit, predictable discrete prior for fine-grained refinement. Specifically, a coarse-to-fine dual-stream Mixture-of-Transformers (MoT) based network is proposed for both dexterous and imagination generation, where high-noise steps establish global interaction dynamics and a contact-aware action expert injects predicted regional contact skill codes at low-noise steps to refine finger articulation and contact transitions. The codes are learned by a spatiotemporal VQ-VAE over six palmar regions and predicted from the instruction and the initial contact state, requiring no future observation at inference. The anatomical contact atlas engine further extracts vertex-level contacts from egocentric videos into six anatomically defined regional atlases. The detailed experiments are conducted to show the effectiveness of our method. On HoloAssist and ARCTIC, CAD-WAM reduces MPJPE by over 39% and wrist translation error by over 30% relative to the strongest baseline. Code will be released.

1 INTRODUCTION

Humans can readily picture how their hands should move to follow an instruction in the scene in front of them, whether using a tool, manipulating an object, assembling parts, or carrying out a lab protocol step by step. Endowing machines with this ability, *i.e.*, generating temporally coherent 4D hand interaction trajectories from an initial egocentric image and a textual instruction, would enable procedural demonstrations at scale. Such hand interaction skills are also a fundamental capability for embodied agents operating in daily environments (Xie et al., 2026; Ye et al., 2026a; Yu et al., 2026; Thambiraja et al., 2026).

Existing 4D hand interaction trajectory methods primarily generate hand motions from egocentric visual observations and language instructions. Early approaches (Prakash et al., 2025) employ Transformer- or diffusion-based predictors to synthesize hand trajectories without any foundational

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.

model prior, limiting their use in real-world scenarios. Inspired by the development of Vision-Language Models (VLMs), Boote et al. (2026); Zhou et al. (2025) leverage pretrained VLMs to provide high-level semantic reasoning for motion generation. However, trajectory generation relies heavily on correlations learned from the training data, and VLMs generalize poorly to unseen interactions and cross-embodiment settings. More recently, Donk (Wang et al., 2026) extends a pretrained video diffusion model with action interfaces for joint video–action denoising, while largely preserving the original visual generation pathway. However, this formulation relies primarily on a generic pretrained video prior with limited task-specific adaptation, rather than explicitly specializing visual imagination and dexterous action modeling for fine-grained manipulation. In addition, existing formulations provide limited reasoning over the fine-grained interaction dynamics that orchestrate how individual fingers engage with the object and remain coordinated with global hand motion over time.

To address these limitations, a MoT-based world action model employs dedicated experts to separately model future visual imagination and dexterous actions, so that physical interaction dynamics can guide hand trajectory generation. Regional contact evolution can serve as an intermediate structure between interaction intent and manipulation skill. This structure describes when and where individual anatomical hand regions establish, maintain, or release contact as the interaction unfolds. An explicit predictive prior over these regional transitions can be inferred from the initial interaction context to provide structured guidance for fine-grained hand interaction skill.

We thus present the *Contact-Aware Dexterous World Action Model* (CAD-WAM), which integrates future video imagination for modeling physical interaction dynamics with reusable contact priors for dexterous trajectory refinement. Specifically, future video imagination is coupled with hand trajectory generation through joint video–action flow matching to model how hand motion evolves within the unfolding interaction. Regional contact priors complement this visual context with explicit guidance on local contact transitions and finger coordination. To make such guidance available from the initial interaction context, *Regional Contact Prior Learning* employs a spatiotemporal VQ-VAE to represent contact evolution across six anatomical regions as discrete contact codes. Future regional code indices are predicted from the language instruction and the initial contact atlas, without access to future contact observations at inference. *Hierarchical Contact-Aware World Action Modeling* integrates the global interaction context and regional contact guidance within a coarse-to-fine MoT architecture. The high-noise regime establishes global interaction semantics, object-relative hand motion, and coarse temporal structure. The low-noise regime then activates a dedicated contact-aware action expert that incorporates the predicted regional contact priors to refine local joint coordination and contact transitions in accordance with the global motion structure. The *Asymmetric Cross-Stream Auxiliary Attention Mask* (ACAAM) enables auxiliary video-flow supervision to guide the contact-aware action refinement expert through asymmetric cross-stream attention, while preventing the refinement stream from accessing future-video tokens. In addition, the *Anatomical Contact Atlas Engine* (ACAE) is developed to extract vertex-level hand object contacts from egocentric videos, partition the MANO (Romero et al., 2017) surface into six palmar regions, and project the regional contacts onto region-wise PCA atlases. Overall, the main contributions are as follows:

- We propose *Contact-Aware Dexterous World Action Model* (CAD-WAM), a coarse-to-fine world action model that jointly performs video–action flow matching, where the high-noise regime establishes global interaction dynamics, object-relative hand motion, and coarse temporal structure, while the low-noise regime activates a dedicated contact-aware action expert to refine fine-grained articulation and contact transitions.
- For contact transition, we propose *Regional Contact Prior Learning* to discretize the per-region contact evolution into a spatiotemporal VQ-VAE codebook and train a conditional indexer predicts the future contact codes from the egocentric observation and the instruction alone, removing the need for any future contact observation at inference.
- In the joint prediction of video and action, *Asymmetric Cross-Stream Auxiliary Attention Mask (ACAAM)* is proposed to propagate auxiliary supervision from future-video modeling to low-noise contact-action refinement while preventing future-video tokens from conditioning hand interaction trajectory generation.
- To map vertex-level global contacts onto local regional contact, *Anatomical Contact Atlas Engine* (ACAE) is developed as the general data process pipeline. Comprehensive quantitative and qualitative evaluations demonstrate that CAD-WAM outperforms state-of-the-art methods in terms of hand translation and joint rotation errors.

2 Related Work

Hand Object Interaction Modeling. Existing hand object interaction modeling can be broadly categorized into object-centric and hand-centric approaches. Object-centric methods condition interaction generation on explicit object geometry. Text2HOI (Cha et al., 2024) predicts a geometry-aware 3D contact map from text and an object mesh, which serves as a strong prior for diffusion-based interaction generation. DiffH2O (Christen et al., 2024) instead decomposes interaction into grasping and manipulation stages, using stage-specific diffusion models to synthesize physically plausible hand motions. Although explicit object representations provide strong geometric constraints, their dependence on specific 3D assets limits applicability in scenarios where 3D object models are unavailable. Hand-centric methods prioritize articulated representations and spatial relationships. GRIP (Taheri et al., 2024) exploits body–object cues to synthesize temporally consistent hand motion, then refines the pose to reduce hand–object penetrations. G-HOP (Ye et al., 2024) represents the hand with a skeletal distance field aligned with the object's latent signed distance field, enabling joint hand–object interaction modeling within a unified diffusion representation. Unlike these studies, CAD-WAM does not require 3D object models at inference and implicitly represents future contact evolution through six anatomically localized contact skill codes.

Hand Object Interaction Generation. LatentAct (Prakash et al., 2025) generates future hand interaction trajectories from a single image, a textual instruction, and contact volume. This design alleviates the constraints that costly 3D asset acquisition imposes on dataset scalability and method applicability. However, LatentAct remains limited in modeling fine-grained and long-range interaction, potentially resulting in insufficient motion diversity and temporal coherence. MEgoHand (Zhou et al., 2025) introduces a vision-language model to infer motion priors and perform spatial reasoning. The absence of contact-aware modeling limits the capability of MEgoHand to capture fine-grained hand object interaction. Different from LatentAct, which jointly tokenizes hand motion and contact maps with the multi-layer residual VQ-VAE (Guo et al., 2024), CAD-WAM separates contact evolution from global hand motion and discretizes regional contact evolution into anatomically structured contact skill codes through a spatiotemporal VQ-VAE. After coarse world action modeling, contact skill codes serve as explicit priors for low-noise action refinement.

World Action Model. Recent world action models (Ye et al., 2026b; Yuan et al., 2026b; Agarwal et al., 2026; Li et al., 2026) jointly generate actions and imagine future manipulation observations. DreamZero (Ye et al., 2026b) adapts a pretrained video diffusion backbone for joint video–action generation and learns from heterogeneous robot trajectories to support zero-shot generalization to unseen tasks. Furthermore, tactile perception introduces physical contact into manipulation skills. OmniVTA (Zheng et al., 2026) generates future visual and tactile dynamics to provide contact guidance for action generation, followed by high-frequency tactile refinement that performs real-time closed-loop correction. VTAM (Yuan et al., 2026a) integrates tactile streams into a pretrained video backbone and introduces deformation-derived virtual-force supervision for action prediction. However, existing methods typically couple contact with motion generation rather than modeling contact evolution across anatomical hand regions. To the best of our knowledge, CAD-WAM is the first contact-aware dexterous world action model tailored to egocentric 4D hand interaction trajectory generation, where anatomical regional contact evolution is formulated as a reusable predicted discrete prior without requiring haptic measurements at inference.

3 Method

3.1 Problem Formulation

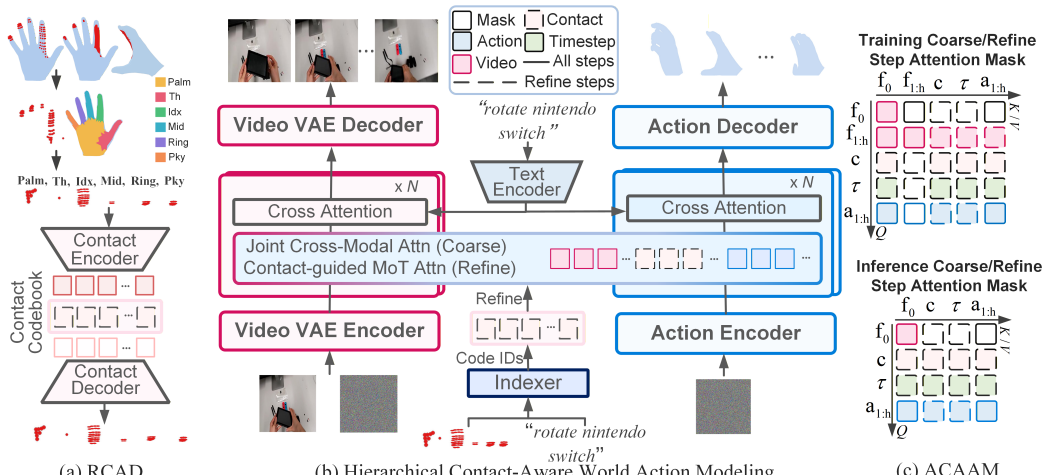
Our task is defined as generating 4D hand interaction trajectory from a textual instruction, an initial RGB frame, and a first-frame palmar contact atlas. Specifically, given a textual instruction ℓ, an initial RGB frame $\mathbf{I}_0$, and a first-frame palmar contact atlas $\mathbf{C}_0$, CAD-WAM, denoted as f_Θ with learnable parameters Θ, generates future egocentric video and the hand interaction trajectory:

$$f_\Theta(\ell, \mathbf{I}_0, \mathbf{C}_0) \longrightarrow \left(\widehat{\mathbf{I}}_{1:F}, \widehat{\mathbf{A}}_{1:H}\right), \tag{1}$$

where $\widehat{\mathbf{I}}_{1:F}$ denotes F generated future RGB frames, and $\mathbf{A}_{1:H} = [\mathbf{a}_1, \ldots, \mathbf{a}_H] \in \mathbb{R}^{H\times 99}$ is a sequence of H per-frame hand states over the same interaction interval. The state $\mathbf{a}_t$ at frame $t \in \{1, \ldots, H\}$ is defined as:

$$\mathbf{a}_t = [\mathbf{r}_t^{\text{hand}}, \mathbf{r}_t^{\text{global}}, \Delta\mathbf{p}_t] \in \mathbb{R}^{99}, \tag{2}$$

where $\mathbf{r}_t^{\text{hand}} \in \mathbb{R}^{90}$, $\mathbf{r}_t^{\text{global}} \in \mathbb{R}^6$, and $\Delta\mathbf{p}_t \in \mathbb{R}^3$ denote the finger rotations, global orientation, and translation offset, respectively.



(a) RCAD (b) Hierarchical Contact-Aware World Action Modeling (c) ACAAM

Figure 2: **CAD-WAM architecture.** (a): RCAD learns a discrete codebook of six region-wise contact evolutions by VQ-VAE. (b): The indexer is trained to infer the code IDs from instruction and the initial contact. Hierarchical contact-aware world action modeling performs coarse denoising at high noise levels and switches to contact-guided MoT refinement at low noise levels. (c) *Top*: The training mask retains future-video tokens for auxiliary supervision but prevents contact and action queries from attending to them. (c) *Bottom*: At inference, future-video tokens are removed, and action queries attend to the observed first-frame, contact map, timestep, and action tokens.

As shown in Figure 2, CAD-WAM comprises *Regional Contact Prior Learning* (Sec. 3.2) and *Hierarchical Contact-Aware World Action Modeling* (Sec. 3.3). The regional contact prior learning first learns a compact codebook of regional contact skills by discretizing contact sequences with a spatiotemporal VQ-VAE. The resulting code indices supervise a conditional contact prior indexer that predicts the corresponding indices from the instruction ℓ and the initial contact atlas $\mathbf{C}_0$. At inference, the predicted indices are used to retrieve embeddings from the frozen codebook, providing explicit contact priors without requiring future contact observations. The hierarchical contact-aware world action modeling is subsequently trained to jointly generate future egocentric video and the hand trajectory $\widehat{\mathbf{A}}_{1:H}$ through coarse-to-fine flow matching with MoT. High-noise steps generate global interaction dynamics, whereas low-noise steps incorporate the contact priors to refine hand articulation and contact transitions.

3.2 Regional Contact Prior Learning

To inject anatomically grounded interaction structure into the world action model, we formulate future regional contact evolution as a discrete predictive interface between semantic interaction intent and dexterous manipulation generation. Dense contact trajectories entangle transient local variations with recurring regional interaction patterns, making direct prediction both high-dimensional and ambiguous. We therefore develop *Regional Contact Prior Learning* (RCPL) to compress these structured dynamics into a compact, reusable, and indexable latent representation. RCPL consists of *Regional Contact Atlas Discretization* (RCAD) and the *Conditional Contact Prior Indexer*. First, the RCAD employs a spatiotemporal VQ-VAE to tokenize regional contact evolution into a shared codebook of reusable contact primitives. Second, the *Conditional Contact Prior Indexer* predicts the future code indices from the egocentric observation and language instruction, allowing region-wise contact transitions to be anticipated without access to future contact observations at inference.

Regional Contact Atlas Discretization. To elevate regional contact evolution from a motion consequence to an explicit and reusable dexterous skill prior, RCAD compresses contact evolution into a discrete codebook (contact skill codes). Learned exclusively from contact maps, the contact skill codes capture recurring patterns of contact evolution across anatomically defined hand regions. The contact evolution is represented as $\mathbf{C} \in \{0,1\}^{N\times R\times P\times P}$. where N, R, and $P \times P$ are the number of temporal frames, the number of anatomical hand regions, and the spatial resolution of each regional contact atlas. The six palmar regions correspond to the Palm, Thumb (**Th**), Index (**Idx**), Middle (**Mid**), Ring, and Pinky (**Pky**).

For each palmar region, the spatial encoder maps the contact atlas into frame-level features. The region-wise contact token $\mathbf{h}_{t,r}$ combines the spatial feature with temporal and regional embeddings that encode frame order and anatomical identity. The shared temporal Transformer contextualizes

the tokens within each region, and temporal mean pooling produces latent $\mathbf{z}_r$ that summarizes the complete regional contact evolution:

$$\begin{aligned} \mathbf{h}_{t,r} &= \phi_{\text{spa}}(\mathbf{C}_{t,r}) + \mathbf{p}_t^t + \mathbf{p}_r^r, \\ \mathbf{z}_r &= \phi_{\text{out}}\left(\text{Mean}_t[\text{TrEnc}(\mathbf{h}_{1:N,r})]\right). \end{aligned} \tag{3}$$

The discrete codebook (contact skill codes) $\mathcal{E} = \{\mathbf{e}_k\}_{k=1}^K$ quantizes each regional feature:

$$\begin{aligned} k_r^* &= \arg\min_k \|\mathbf{z}_r - \mathbf{e}_k\|_2^2, \\ \overline{\mathbf{z}}_r &= \mathbf{z}_r + \text{sg}\left(\mathbf{e}_{k_r^*} - \mathbf{z}_r\right), \end{aligned} \tag{4}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The Transformer decoder reconstructs the complete contact sequence using learned time region queries and the quantized regional embeddings as memory. Given the reconstructed contact probabilities $\widehat{\mathbf{C}}$, the optimization objective of RCAD is defined as:

$$\mathcal{L}_{\text{VQ}} = \mathcal{L}_{\text{WBCE}}(\widehat{\mathbf{C}}, \mathbf{C}) + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}(\widehat{\mathbf{C}}, \mathbf{C}) + \beta\frac{1}{R}\sum_{r=1}^{R}\|\mathbf{z}_r - \text{sg}(\mathbf{e}_{k_r^*})\|_2^2, \tag{5}$$

where $\mathcal{L}_{\text{WBCE}}$, λ_{dice}, and β are 8.0, 0.5, and 0.25, respectively. The contact skill codes are updated using exponential moving averages of the assigned encoder latents. After training, RCAD provides a frozen codebook and represents each contact using six discrete indices $\mathbf{k}^* = (k_1^*, \ldots, k_R^*)$, where k_r^* denotes the codebook entry assigned to the r-th palmar region.

Conditional Contact Prior Indexer. To predict contact evolution codes $\mathbf{k}^*$ at inference, the conditional contact prior indexer takes the language instruction ℓ and initial contact atlas $\mathbf{C}_0$ as inputs. During training, contact skill code supervision is obtained by encoding complete training windows with the frozen RCAD encoder. Future contact information is used solely for label construction and is never provided to the conditional contact before the indexer during inference.

Frozen T5 features $\mathbf{T}_\ell$ are projected into a d-dimensional token space. The flattened first-frame regional atlas is projected into one contact token. The projected language tokens and the contact token are concatenated to form the multimodal memory $\mathbf{M} = [\phi_{\text{text}}(\mathbf{T}_\ell); \phi_{\text{fc}}(\text{vec}(\mathbf{C}_0))]$, where $[\cdot;\cdot]$ denotes token-wise concatenation. The two-layer Transformer decoder takes six learned region queries as queries and $\mathbf{M}$ as the key-value memory. The decoder outputs region-wise representations for contact skill code prediction. For each region r, the region-wise classifier produces logits $\boldsymbol{\pi}_r \in \mathbb{R}^K$, and the predicted contact skill code is $\widehat{k}_r = \arg\max_k \boldsymbol{\pi}_{r,k}$. The optimization objective of the conditional contact prior indexer is defined as:

$$\mathcal{L}_{\text{idx}} = -\frac{1}{R}\sum_{r=1}^{R}\log \text{Softmax}(\boldsymbol{\pi}_r)_{k_r^*}. \tag{6}$$

The predicted indices retrieve embeddings from the frozen contact skill codes. The learned region embedding $\boldsymbol{\rho}_r$ is added before projecting each embedding into a contact token. The six contact skill codes $\{\mathbf{c}_r\}_{r=1}^R$ are obtained as $\mathbf{c}_r = \phi_{\text{code}}(\mathbf{e}_{\widehat{k}_r} + \boldsymbol{\rho}_r)$.

3.3 Hierarchical Contact-Aware World Action Modeling

As shown in Figure 2, to jointly model physical hand interaction dynamics and contact evolution, hierarchical contact-aware world action modeling couples video world modeling with hand interaction trajectory generation through a noise-dependent coarse-to-fine architecture. The dual stream video–action flow matching establishes global interaction dynamics at high noise levels, after which predicted regional contact priors guide fine-grained hand motion refinement at low noise levels. The hierarchical contact-aware world action modeling takes an egocentric image, the first-frame contact state, and a textual instruction as input. The output is 4D hand interaction trajectory comprising hand poses and global positions over time steps. The contact skill codes learned by regional contact atlas discretization and conditional contact prior indexer are frozen. Given the textual instruction and the initial contact atlas, the indexer predicts six regional code indices that identify the corresponding embeddings in the frozen contact skill codes. Hierarchical contact-aware world action modeling then injects the embeddings as contact skill codes into low-noise motion refinement.

Dual Stream World–Action Backbone. The dual stream Mixture-of-Transformers (Liang et al., 2025) architecture comprises a video expert initialized from Wan (Wan et al., 2025) and an action DiT expert (Yuan et al., 2026b). For modality $m \in \{\mathrm{vid}, \mathrm{act}\}$, the noisy state and the corresponding target velocity are defined as:

$$\mathbf{x}_\sigma^m = (1-\sigma)\mathbf{x}_0^m + \sigma\boldsymbol{\epsilon}^m, \qquad \mathbf{v}^{m,*} = \boldsymbol{\epsilon}^m - \mathbf{x}_0^m, \tag{7}$$

where $\mathbf{x}_0^m$ denotes the clean modality state, $\boldsymbol{\epsilon}^m$ denotes Gaussian noise, and σ specifies the noise level. The flow matching objective is defined as:

$$\mathcal{L}_{\mathrm{FM}}^m = \mathbb{E}_{\sigma,\boldsymbol{\epsilon}}\left[w(\sigma)\left\|\mathbf{v}_\theta^m\left(\mathbf{x}_\sigma^m\right) - \mathbf{v}^{m,*}\right\|_2^2\right]. \tag{8}$$

Textual instruction and first-frame observation are omitted from $\mathbf{v}_\theta^m$ for notational clarity. The coarse action sequence is defined as:

$$\mathbf{X}_{\mathrm{coarse}} = \left[\phi_{\mathrm{act}}(\mathbf{A}_\sigma)\right], \tag{9}$$

where ϕ_{act} denotes the action token embedding of the action DiT.

Noise-Dependent Coarse-to-Fine Modeling. At high noise levels, the video expert and the base action expert perform dual stream flow matching to generate global interaction dynamics and coarse hand trajectory. Regional contact atlas tokens are reserved for the low noise stage, where fine-grained hand articulation and contact transitions become recoverable. $\sigma \le \sigma_{\mathrm{ref}}$, action denoising switches from the base action expert to a separately parameterized contact-action expert, while the video expert remains shared across the two noise regimes. During training, the base video–action flow matching objectives remain active, while an additional low-noise refinement objective trains the contact-action expert. The regional indices predicted by the Conditional Contact Prior Indexer retrieve the corresponding embeddings from the frozen contact skill codes. The retrieved embeddings form the regional contact skill codes $\mathbf{c}_1, \ldots, \mathbf{c}_R$. The low-noise contact-action expert processes these tokens together with an implicit noise-level token and the noisy future-action tokens:

$$\mathbf{X}_{\mathrm{ref}} = \left[\mathbf{c}_1, \ldots, \mathbf{c}_R; \boldsymbol{\tau}_\sigma; \phi_{\mathrm{act}}(\mathbf{A}_\sigma)\right], \tag{10}$$

where $\boldsymbol{\tau}_\sigma$ denotes the implicit noise-level token. During refinement, the regional contact skill codes, noise-level token, and noisy future-action tokens interact through bidirectional self-attention while retaining language cross-attention and observed first-frame video conditioning. Only the future-action suffix associated with $\phi_{\mathrm{act}}(\mathbf{A}_\sigma)$ is decoded into action velocity. The contact and noise-level prefixes are removed before decoding and loss computation and therefore serve as fixed conditioning signals rather than denoising targets.

Asymmetric Cross-Stream Auxiliary Attention Mask. As shown in Figure 2 (c), ACAAM introduces asymmetric attention between future-video tokens and the refinement sequence $\mathbf{X}_{\mathrm{ref}}$ defined above. Denote the observed first-frame and future-video tokens by $\mathbf{f}_0$ and $\mathbf{f}_{1:F}$, respectively. Future-video queries can therefore access refinement representations, while refinement queries cannot access future-video tokens. Future-video queries therefore access refinement representations, and $\mathcal{L}_{\mathrm{vid}}^{\mathrm{aux}}$ propagates gradients through the key–value representations supplied by the refinement expert. Refinement queries attend bidirectionally within $\mathbf{X}_{\mathrm{ref}}$ and access only $\mathbf{f}_0$ as visual context. First-frame video queries attend exclusively to first-frame video keys, which also prevents indirect access to future-video information across layers. At inference, future-video tokens are omitted by default; when video is decoded for visualization, they remain masked from the refinement stream, while refinement retains the first-frame visual context and predicted regional contact prior.

Training Loss. The optimization objective of hierarchical contact-aware world action modeling is defined as:

$$\mathcal{L}_{\mathrm{CAD}} = \lambda_{\mathrm{vid}}\mathcal{L}_{\mathrm{FM}}^{\mathrm{vid}} + \lambda_{\mathrm{act}}\mathcal{L}_{\mathrm{FM}}^{\mathrm{act}} + \lambda_{\mathrm{refine}}\mathcal{L}_{\mathrm{refine}} + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}} + \lambda_{\mathrm{vaux}}\mathcal{L}_{\mathrm{vid}}^{\mathrm{aux}}, \tag{11}$$

where λ_{vid}, λ_{act}, $\lambda_{\mathrm{refine}}$, λ_{ref}, and λ_{vaux} are set to 1, 1, 1, 0.1, and 0.05, respectively. The video and action experts are supervised by $\mathcal{L}_{\mathrm{FM}}^{\mathrm{vid}}$ and $\mathcal{L}_{\mathrm{FM}}^{\mathrm{act}}$, respectively. The low-noise contact-action expert is supervised by an additional action flow matching loss $\mathcal{L}_{\mathrm{refine}}$. For reference-space supervision, the clean action estimate is recovered from the base action branch as $\widehat{\mathbf{A}}^{\,\mathrm{clean}} = \mathbf{A}_\sigma - \sigma\mathbf{v}_\theta^{\mathrm{act}}$. After denormalization, the clean estimate is converted into reference-frame joints as $\widehat{\mathbf{J}} = \mathcal{F}_{\mathrm{MANO}}(\widehat{\mathbf{A}}^{\,\mathrm{clean}})$, where $\mathcal{F}_{\mathrm{MANO}}$ includes MANO reconstruction and the subsequent transformation into the reference coordinate system. The reference-space loss is defined as $\mathcal{L}_{\mathrm{ref}} = \mathbb{E}_{t,j}\|\widehat{\mathbf{J}}_{t,j} - \mathbf{J}_{t,j}^*\|_1$.

4 EXPERIMENTS

4.1 EVALUATION METRICS

The model is evaluated using Mean Per Joint Position Error (MPJPE), Mean Per Vertex Error (MPVE), Procrustes-Aligned MPJPE (PA-MPJPE), Procrustes-Aligned MPVE (PA-MPVE), Mean Wrist Translation Error (MWTE), and Mean Rotation Error (MRE). Specifically, MPJPE and MPVE measure the mean Euclidean errors between predicted and ground-truth MANO joint positions and mesh vertices, respectively. PA-MPJPE and PA-MPVE measure the corresponding errors after Procrustes alignment, which removes global differences in scale, rotation, and translation, and therefore isolate the accuracy of hand articulation from that of global hand placement. MWTE measures the mean Euclidean error between predicted and ground-truth wrist translations. MRE measures the mean geodesic rotation error over the wrist and finger joints. MPJPE, PA-MPJPE, MPVE, PA-MPVE, and MWTE are reported in centimeters, whereas MRE is reported in radians. More details are provided in Appendix A.4.

4.2 DATASET PREPROCESSING

Anatomical Contact Atlas Engine. As shown in Figure 3, the *Anatomical Contact Atlas Engine* (ACAE) is developed to construct anatomically organized regional contact representations. ACAE transforms vertex-level pseudo-contact cues into spatially aligned contact atlases over six palmar regions, providing structured contact representations for *regional contact prior learning*. Object masks and hand estimates, including MANO poses and rendered 2D hand masks, are obtained from SAM2 (Ravi et al., 2025) and HaMeR (Pavlakos et al., 2024), respectively. Pseudo-contact labels are then estimated by projecting the reconstructed MANO vertices onto the image plane and identifying vertices whose projections fall within the 2D hand–object overlap regions. Given the vertex-level pseudo-contact labels, ACAE reorganizes the contact representation according to hand anatomy and local surface geometry. The 778 MANO vertices are first partitioned into six palmar regions: Palm, Thumb, Index, Middle, Ring, and Pinky. A global palm-plane projection establishes a consistent hand-centric reference frame, while a region-wise PCA projection parameterizes each anatomical region in the 2D space to reduce the geometric distortion caused by flattening the irregular palmar surface. The projected contact vertices are rasterized into six spatially aligned 32×32 regional contact atlases. This construction preserves anatomical correspondence and local spatiotemporal contact structure, and provides structured regional contact representations for subsequent contact discretization and prior prediction without introducing additional contact inference. The main experiments are conducted on HoloAssist (Wang et al., 2023) and ARCTIC (Fan et al., 2023) datasets. Ablation studies are conducted on HoloAssist. Dataset statistics and splits are provided in Appendix A.2 and Appendix A.8, respectively.

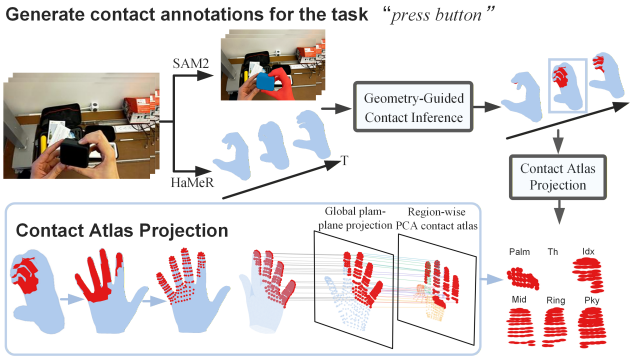


Figure 3: **Anatomical Contact Atlas Engine.** Object masks and 3D MANO hand are first processed by geometry-guided contact inference to obtain pseudo-contact maps. The inferred contacts are then transformed through contact atlas projection into localized regional contact atlases.

4.3 COMPARISON WITH STATE-OF-THE-ART METHODS

CAD-WAM is compared with state-of-the-art methods, including: **LatentAct** (Prakash et al., 2025) employs a discrete trajectory codebook and a Transformer-based generative decoder to generate future hand motions and contact maps from textual languages, initial observations, and contact volumes. **LatentAct-Diff** (Prakash et al., 2025) adopts a motion diffusion model (Tevet et al., 2023) as generative decoder. **MEgoHand** (Zhou et al., 2025) (UniDepth V2 (Piccinelli et al., 2025)), a strong recent VLA-style baseline, combines VLM motion reasoning and monocular depth perception with a DiT-based flow-matching policy. For all the baselines, we follow the official implementations and retrain these methods under the same dataset splits.

As shown in Table 1, CAD-WAM achieves the best performance for all six metrics on HoloAssist and ARCTIC. With the lowest baseline value for each metric as the reference, CAD-WAM reduces MPJPE by 39.3% on HoloAssist and 44.9% on ARCTIC. CAD-WAM consistently improves absolute hand translation, articulated pose, and joint rotation accuracy across the benchmarks. Results on

Table 1: Quantitative comparison on the HoloAssist and ARCTIC datasets. CAD-WAM outperforms the state-of-the-art methods.

Dataset	Method	MPJPE↓	PA-MPJPE↓	MPVE↓	PA-MPVE↓	MWTE↓	MRE↓
HoloAssist	LatentAct (Prakash et al., 2025)	15.09	1.17	15.00	1.11	14.36	0.40
	– no contact map	13.35	1.11	13.28	1.05	12.58	0.35
	LatentAct-Diff (Prakash et al., 2025)	17.87	1.03	17.76	0.98	17.34	0.38
	– no contact map	17.29	1.05	17.23	1.00	16.62	0.33
	MEgoHand (Zhou et al., 2025)	14.61	1.89	14.37	1.84	11.21	0.62
	CAD-WAM	**8.10**	**0.86**	**8.06**	**0.82**	**7.81**	**0.25**
ARCTIC	LatentAct (Prakash et al., 2025)	11.93	1.48	11.84	1.43	10.78	0.63
	– no contact map	12.88	1.43	12.83	1.40	12.35	0.66
	LatentAct-Diff (Prakash et al., 2025)	14.87	1.54	14.83	1.48	14.19	0.71
	– no contact map	13.25	1.60	13.18	1.56	12.22	0.69
	MEgoHand (Zhou et al., 2025)	11.65	1.48	11.60	1.45	11.00	0.57
	CAD-WAM	**6.42**	**1.16**	**6.38**	**1.13**	**5.70**	**0.44**

test splits involving unseen interaction combinations and cross-subject evaluation demonstrate the generalization capability of CAD-WAM. The results indicate the effectiveness of the contact evolution modeling and physical interaction priors introduced in CAD-WAM for contact-aware hand trajectory generation. We also illustrate the visualized results of different methods by projected the trajectories onto the corresponding RGB frames, As shown in Figure 4, CAD-WAM outperforms state-of-the-art methods in terms of wrist trajectory and finger articulation accuracy, while the qualitative overlays indicate improved hand alignment throughout the interaction sequence. Additional qualitative results are provided in Appendix A.5.

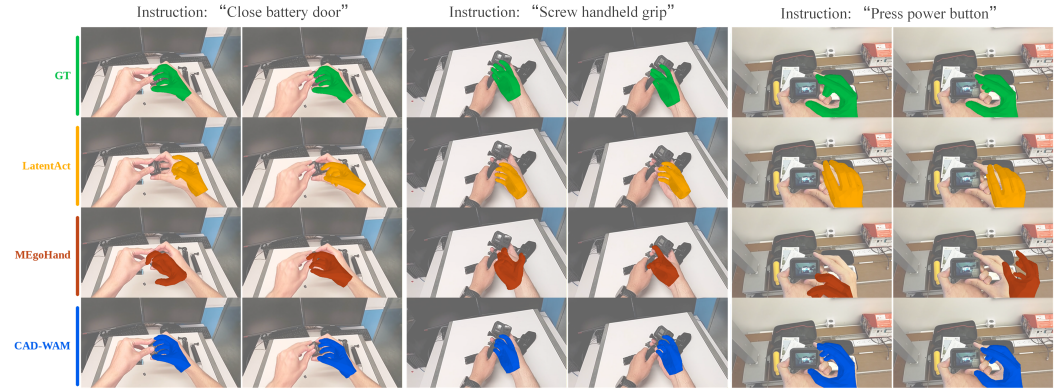


Figure 4: Qualitative comparisons. CAD-WAM generates accurate wrist motion, finger articulation, and hand object geometric alignment.

4.4 Ablation Study

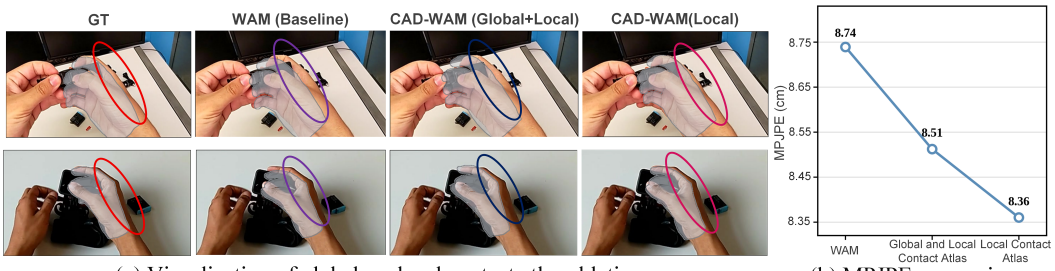
The ablation studies examine several key design choices, including the contact representation, codebook capacity, global vs. local contact atlases for contact-aware action refinement, and the ACAAM. The underlying dual stream MoT video-action model before these designs are introduced is denoted as WAM (Baseline). As shown in Table 2, CAD-WAM with contact-aware action refinement (image contact skill codes, local contact atlas, and a codebook capacity of 64) and the ACAAM achieves the best performance. These consistent increases in MPJPE, MPVE, and MWTE further indicate that CAD-WAM benefits from imagining future physical interaction dynamics, while explicit contact priors provide additional guidance for hand trajectory refinement. Additional ablation studies are provided in Appendix A.9.

Table 2: Ablation study on the HoloAssist Dataset.

Method	MPJPE↓	MPVE↓	MWTE↓
WAM (Baseline)	8.74	8.71	8.59
w/ vertex contact skill codes	8.53	8.51	8.35
w/ image contact skill codes	8.36	8.33	8.18
CAD-WAM (Full Method)	**8.10**	**8.06**	**7.81**

Global vs. Local Contact Atlas. Since global layouts are commonly used for contact representation, the global contact atlas is included for comparison. The global atlas integrates six region-wise PCA projections into a unified whole-hand layout with preserved relative positions, while the local atlases retain six independent palmar regions for the palm and five fingers. As shown in Figure 5, the local contact atlas, comprising six palmar contact region tokens, achieves MPJPE of 8.36 cm. RCAD suffers from codebook collapse when trained solely on the global whole-palm contact atlas. CAD-WAM trained with the global whole-palm contact atlas and the six regional contact atlases

achieves MPJPE of 8.51 cm. The degradation indicates that the global token provides no additional benefit. The results suggest that anatomical regional decomposition preserves localized spatiotemporal variations in contact dynamics effectively.



(a) Visualization of global vs. local contact atlas ablations. (b) MPJPE comparison.

Figure 5: Global vs. local contact atlas ablations. Local contact atlas achieves the best performance.

Contact Representation. Since MANO vertex binary contact states are commonly used for contact representation, vertex contact skill codes are included in the ablation study. As shown in Table 2, refining the action expert with vertex contact skill codes and image contact skill codes reduces MPJPE by 2.35% and 4.34%, respectively. Vertex contact skill codes are discrete latent codes learned from the temporal binary contact states of MANO vertices. Image contact skill codes are obtained by projecting palm-side contact vertices onto six palmar regions and discretizing the spatiotemporal contact maps. The consistent gains across joint and wrist translation accuracy demonstrate that regional contact atlases preserve localized spatiotemporal contact structure effectively and provide a strong prior for contact-aware action refinement.

ACAAM Ablation. As shown in Figure 6, the proposed ACAAM achieves the lowest MPJPE and MPVE of 8.28 cm and 8.24 cm, respectively. Without the refined video tokens, MPJPE and MPVE increase to 8.36 cm and 8.33 cm, which supports the use of auxiliary future-video supervision during contact-aware refinement. The removal of the timestep token τ causes performance degradation, with MPJPE and MPVE increasing to 8.52 cm and 8.49 cm, respectively. The results establish the importance of implicit noise-level conditioning for the low-noise contact-action expert. The additional asymmetric constraint between contact and action tokens increases MPJPE and MPVE to 8.43 cm and 8.39 cm, respectively. Attention mask visualizations for ACAAM ablations are provided in Appendix A.9.

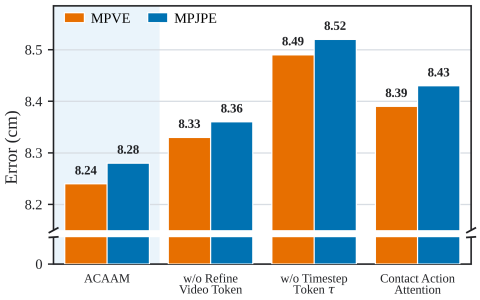


Figure 6: ACAAM ablations.

Limitation. Many everyday objects are thin, small, or slender, and obtaining accurate 3D assets remains challenging. Our study focuses on generating 4D hand interaction trajectories from a single image without relying on 3D geometry. In the absence of 3D object geometry, hand object interpenetration metrics that require such geometry do not need to be evaluated. Contact evolution in the current formulation describes the spatiotemporal changes of hand contact location. Physical haptic properties such as pressure, shear, and slip are not incorporated in this study.

5 Conclusion

We propose CAD-WAM for generating 4D hand interaction trajectories from a single egocentric image and a language instruction. CAD-WAM first captures coarse interaction dynamics at high noise levels, followed by contact-aware hand motion refinement guided by regional contact priors at low noise levels. The proposed Asymmetric Cross-Stream Auxiliary Attention Mask further introduces auxiliary video-flow supervision without exposing action prediction to future visual information. To support regional contact modeling, the ACAE extracts vertex-level hand object contacts from egocentric videos and projects the contacts onto six locally parameterized PCA atlases, producing anatomy-aligned contact evolution. Extensive experiments demonstrate substantial reductions in hand translation and joint rotation errors, validating the effectiveness of regional contact priors for fine-grained action refinement. More reliable in-the-wild contact representation, adaptive hand-surface partitioning, and long-horizon contact modeling could further improve data scalability and generalization to complex dexterous interactions.

6 AI USE STATEMENT

In this work, we used generative AI tools to improve the readability, grammar, and clarity of the manuscript. All AI-assisted outputs were reviewed by the authors, including manual verification of language edits and validation of AI-generated synthetic data before their use in experiments. We take responsibility for the final content of this work, including text, claims, data, and other artifacts produced with the aid of generative AI.

7 REPRODUCIBILITY STATEMENT

The model architecture, training objectives, and the training pipeline of CAD-WAM are described in Section 3, and the dataset preprocessing and evaluation protocol are described in Sections 4.2 and 4.1. In the appendix, Appendix A.2 details the HOI datasets and data splits, Appendix A.3 reports the complete training configuration and hyperparameters (Table 3) together with the inference setting, Appendix A.4 gives the exact definitions of all evaluation metrics, Appendix A.1 describes how the baselines are obtained and evaluated, and Appendix A.6 lists the computational resources. Appendix A.8 further provides statistics of the constructed regional contact data. Code will be made publicly available upon acceptance.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Bikram Boote, Junho Kim, Ozgur Kara, Sangmin Lee, and James M Rehg. Coherenthand: Temporally consistent 3d hand trajectory synthesis with semantic motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3417–3427, 2026.

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.

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.

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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Nu6N69i8SB.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Aditya Prakash, Benjamin Lundell, Dmitry Andreychuk, David Forsyth, Saurabh Gupta, and Harpreet Sawhney. How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7026–7036, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 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), November 2017.

Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J Black. Grip: Generating interaction poses using spatial cues and latent consistency. In *2024 International Conference on 3D Vision (3DV)*, pp. 933–943. IEEE, 2024.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SJ1kSyO2jwu.

Balamurugan Thambiraja, Omid Taheri, Radek Danecek, Giorgio Becherini, Gerard Pons-Moll, and Justus Thies. CLUTCH: Contextualized language model for unlocking text-conditioned hand motion modelling in the wild. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=W7YRskO47j.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Dingrui Wang, YuAn Wang, Jinkun Liu, Yue Zhang, Mattia Piccinini, Yu Sun, and Johannes Betz. Unified video-action joint denoising for dexterous action and data generation. *arXiv preprint arXiv:2606.03868*, 2026.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. 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 (ICCV)*, pp. 20270–20281, October 2023.

Binzhu Xie, Shi Qiu, Sicheng Zhang, Yinqiao Wang, Hao Xu, Muzammal Naseer, Chi-Wing Fu, and Pheng-Ann Heng. Egohandicl: Egocentric 3d hand reconstruction with in-context learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Qi Ye, Qingtao Liu, Siyun Wang, Jiaying Chen, Yu Cui, Ke Jin, Huajin Chen, Xuan Cai, Gaofeng Li, and Jiming Chen. Visual-tactile pretraining and online multitask learning for humanlike manipulation dexterity. *Science Robotics*, 11(110):eady2869, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Naveed Malik, Kyungmin Lee, William Liang, Nadun Ranawaka Arachchige, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Danfei Xu, Yilun Du, Ryan Julian, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Fan, and Joel Jang. World action models are zero-shot policies. In *ICLR 2026 the 2nd Workshop on World Models: Understanding, Modelling and Scaling*, 2026b. URL https://openreview.net/forum?id=cd33uUB609.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1920, 2024.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, et al. Forcevla: Enhancing vla models with a force-aware moe for contact-rich manipulation. *Advances in Neural Information Processing Systems*, 38:93409–93439, 2026.

Haoran Yuan, Weigang Yi, Zhenyu Zhang, Wendi Chen, Yuchen Mo, Jiashi Yin, Xinzhuo Li, Xiangyu Zeng, Chuan Wen, Cewu Lu, et al. Vtam: Video-tactile-action models for complex physical interaction beyond vlas. *arXiv preprint arXiv:2603.23481*, 2026a.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, et al. Omnivta: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

Bohan Zhou, Yi Zhan, Zhongbin Zhang, and Zongqing Lu. Megohand: Multimodal egocentric hand-object interaction motion generation. *Advances in Neural Information Processing Systems*, 38:49464–49490, 2025.

A APPENDIX

A.1 BASELINE METHODS

CAD-WAM is compared with LatentAct[1] (Prakash et al., 2025) and MEgohand[2] (Prakash et al., 2025) for 4D hand interaction trajectory generation. Unlike LatentAct, which filters out valid samples with large hand displacements in training and testing, and MEgoHand, which uses a separate mixed training set and applies such filtering only at test time, CAD-WAM retains these samples for training and evaluation.

A.2 HOI DATASETS

The experiments utilize a variety of publicly available egocentric hand object interaction datasets. Each dataset is introduced below together with a link to its official website.

ARCTIC:[3] A richly annotated dataset for tracking hand object contact and motion in egocentric scenarios.

HOLO:[4] A large-scale dataset of household manipulation tasks captured in real-world environments.

The main and ablation experiments. Specifically, we construct 33-frame motion clips with a temporal stride of 10 and remove windows containing no hand object contact. The training split additionally excludes text prompts associated with fewer than 10 valid contact windows. HOLO and ARCTIC adopt distinct generalization protocols. HoloAssist follows a compositional task split, which includes unseen verb–noun combinations in the test dataset. The training, validation, and test datasets contain 30,654, 3,082, and 4,480 clips, respectively. ARCTIC follows a cross-subject split, where sequences from `s01`, `s02`, `s04`, and `s06-s10` form the training set, while `s05` forms the held-out evaluation set. For Arctic, the training, validation, and test datasets contain 6,402, 510, and 510 clips, respectively. Unlike LatentAct, which filters out valid samples with large hand displacements, all dataset samples are retained for evaluation.

A.3 IMPLEMENTATION DETAILS

The main hyperparameters used in the training for the main experiments are reported in Table 3.

A.4 EVALUATION METRICS

CAD-WAM does not generate future contacts, which predicts regional contact skill codes as latent priors for refining future hand actions. The setting does not assume access to object meshes, CAD models, or other 3D object geometry. Specifically, MPJPE and MPVE measure the mean Euclidean errors between predicted and ground truth MANO joint positions and mesh vertices, respectively. MPJPE-PA and MPVE-PA measure the corresponding errors after Procrustes alignment, which removes global differences in scale, rotation, and translation. MWTE measures the mean Euclidean error between predicted and ground truth wrist translations. MRE measures the mean geodesic rotation error over the wrist and 15 finger joints in radians. We evaluate action generation using six metrics that measure joint translation, and rotation accuracy. Let $\mathcal{Q}$ denote the set of all valid evaluated frames, and let $|\mathcal{Q}|$ be its cardinality. For a frame $q \in \mathcal{Q}$, we denote the predicted and ground truth 3D joint positions by $\widehat{\mathbf{p}}_{q,j}, \mathbf{p}_{q,j} \in \mathbb{R}^3$, and the corresponding MANO vertices by $\widehat{\mathbf{u}}_{q,v}, \mathbf{u}_{q,v} \in \mathbb{R}^3$. We use $J = 21$ hand joints and $V = 778$ MANO mesh vertices.

[1] `https://github.com/ap229997/latentact`
[2] `https://github.com/BeingBeyond/MEgoHand`
[3] `https://arctic.is.tue.mpg.de/`
[4] `https://holoassist.github.io/`

Table 3: Hyperparameters of training pipeline.

Regional Contact Atlas Discretization

Hyperparameter	Value
Atlas resolution	32×32
Hand regions	6
Encoder layers	2
Decoder layers	2
Hidden dimension	256
Attention heads	4
Dropout	0.1
Codebook size	64
Code dimension	256
Codebook sharing	Shared
Training epochs	120
Batch size	64
Learning rate	2×10^{-4}
Optimizer	AdamW
Weight decay	1×10^{-4}
EMA decay	0.99
Commitment weight	0.25
Positive BCE weight	8.0
Dice-loss weight	0.5

Conditional Contact Prior Indexer

Hyperparameter	Value
Transformer layers	2
Hidden dimension	256
Attention heads	4
Dropout	0.1
Predicted code indices	6
Codebook	Frozen
Training objective	Cross-entropy
Maximum epochs	50
Batch size	256
Learning rate	1×10^{-4}
Optimizer	AdamW
Weight decay	1×10^{-4}
LR schedule	Fixed
Early stopping	Enabled
Patience	8 epochs
Minimum improvement	5×10^{-4}

Hierarchical Contact-Aware World Action Modeling

Hyperparameter	Value
Video initialization	Wan2.2-TI2V-5B
Action initialization	ActionDiT
Transformer layers	30
Attention heads	24
Video hidden dimension	3072
Action hidden dimension	1024
Codebook/indexer	Frozen
Training epochs	10
GPUs	2 H100s
Per-GPU batch size	64
Global batch size	128
Optimizer	AdamW
Adam β_1, β_2	0.9, 0.95
Initial learning rate	1×10^{-4}
Final learning rate	1×10^{-6}
Weight decay	0
LR schedule	Cosine
Warm-up ratio	0.05
Training timesteps	1,000
Flow-shift factor	5
Inference steps	10
Expert threshold	$\sigma \leq 0.77$
Flow loss weights	1.0 / 1.0
Reference loss weight	0.1

Mean Per-Joint Position Error (MPJPE). MPJPE measures the average Euclidean distance between predicted and ground truth 3D joint positions:

$$\mathrm{MPJPE} = \frac{1}{|\mathcal{Q}|J} \sum_{q\in\mathcal{Q}} \sum_{j=1}^{J} \|\widehat{\mathbf{p}}_{q,j} - \mathbf{p}_{q,j}\|_2 \,. \tag{12}$$

This metric evaluates absolute joint accuracy and therefore includes errors in global hand translation and orientation.

Procrustes-Aligned MPJPE (PA-MPJPE). For each valid frame, we independently estimate the similarity transformation $\mathcal{S}_q^{\mathrm{joint}}$—including scale, rotation, and translation—that optimally aligns the predicted joints with the ground truth. PA-MPJPE is defined as

$$\text{PA-MPJPE} = \frac{1}{|\mathcal{Q}|J} \sum_{q\in\mathcal{Q}} \sum_{j=1}^{J} \left\|\mathcal{S}_q^{\mathrm{joint}}(\widehat{\mathbf{p}}_{q,j}) - \mathbf{p}_{q,j}\right\|_2 \,. \tag{13}$$

PA-MPJPE removes global similarity-transform discrepancies and primarily measures articulated hand-pose accuracy.

Mean Per-Vertex Error (MPVE). MPVE evaluates the geometric accuracy of the reconstructed MANO surface:

$$\mathrm{MPVE} = \frac{1}{|\mathcal{Q}|V} \sum_{q\in\mathcal{Q}} \sum_{v=1}^{V} \|\widehat{\mathbf{u}}_{q,v} - \mathbf{u}_{q,v}\|_2 \,. \tag{14}$$

Procrustes-Aligned MPVE (PA-MPVE). We independently fit a per-frame similarity transformation $\mathcal{S}_q^{\mathrm{vert}}$ using the complete set of MANO vertices:

$$\text{PA-MPVE} = \frac{1}{|\mathcal{Q}|V} \sum_{q\in\mathcal{Q}} \sum_{v=1}^{V} \left\|\mathcal{S}_q^{\mathrm{vert}}(\widehat{\mathbf{u}}_{q,v}) - \mathbf{u}_{q,v}\right\|_2 \,. \tag{15}$$

This metric measures local mesh deformation accuracy after removing global scale, rotation, and translation differences.

Mean Wrist Translation Error (MWTE). Let $\widehat{\mathbf{d}}_q, \mathbf{d}_q \in \mathbb{R}^3$ denote the predicted and ground truth absolute wrist translations in the camera coordinate system. MWTE is defined as:

$$\mathrm{MWTE} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left\| \widehat{\mathbf{d}}_q - \mathbf{d}_q \right\|_2 . \tag{16}$$

It isolates the global translation accuracy of the generated hand trajectory.

Mean Rotation Error (MRE). Let $\widehat{\mathbf{G}}_{q,k}, \mathbf{G}_{q,k} \in \mathrm{SO}(3)$ denote the predicted and ground truth rotation matrices for joint k. We evaluate $K = 16$ rotations, comprising one global wrist rotation and 15 articulated finger-joint rotations. MRE is the average geodesic distance on $\mathrm{SO}(3)$:

$$\mathrm{MRE} = \frac{1}{|\mathcal{Q}|K} \sum_{q \in \mathcal{Q}} \sum_{k=1}^{K} \cos^{-1} \left(\mathrm{clip} \left(\frac{\mathrm{tr}\left(\mathbf{G}_{q,k}^{\top} \widehat{\mathbf{G}}_{q,k} \right) - 1}{2}, -1, 1 \right) \right) . \tag{17}$$

MRE lies in $[0, \pi]$ radians, where a lower value indicates more accurate joint rotations. MPJPE, PA-MPJPE, MPVE, PA-MPVE, and MWTE are reported in centimeters, while MRE is reported in radians.

A.5 More Qualitative Results

As shown in Figure 7, additional qualitative comparisons of CAD-WAM, MEgoHand, and LatentAct are provided.

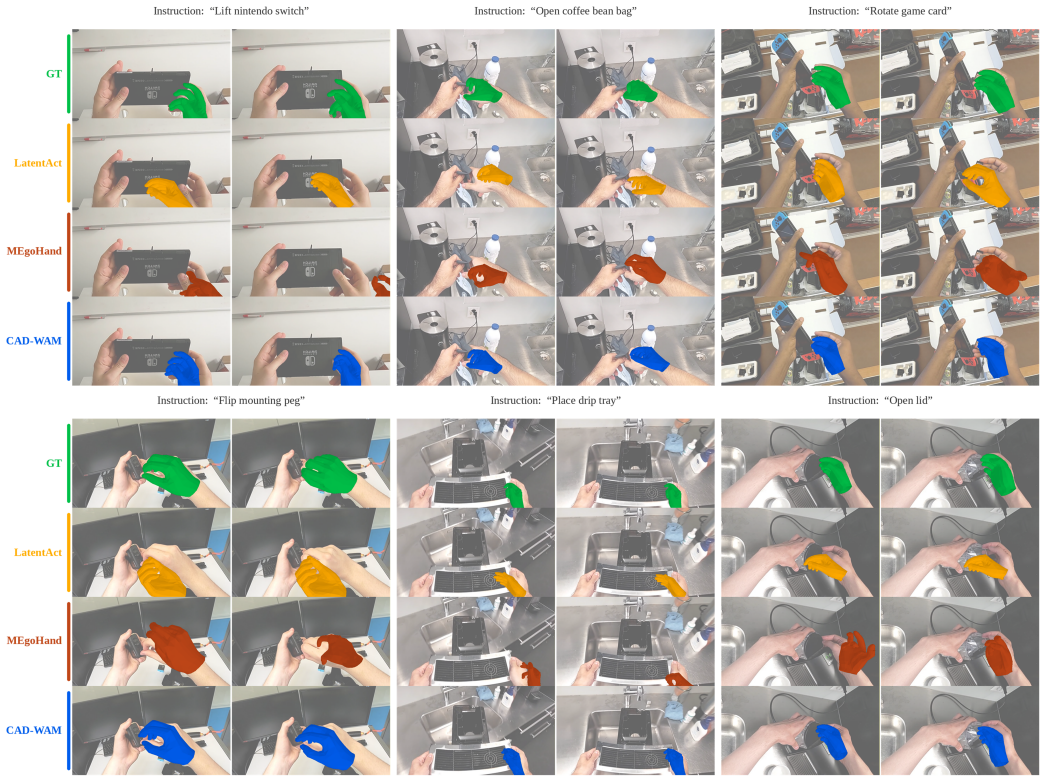


Figure 7: More Qualitative comparisons of CAD-WAM, MEgoHand, and LatentAct.

A.6 Computational Resources.

The regional contact atlas discretization of CAD-WAM are trained on a single 80 GB NVIDIA H100 GPU, while the hierarchical contact-aware world action modeling is trained using two 80 GB NVIDIA H100 GPUs with DeepSpeed ZeRO-2. The evaluation is performed on two H100 GPUs.

A.7 More Contact Visualizations

As shown in Figure 8, more contact visualizations are provided.

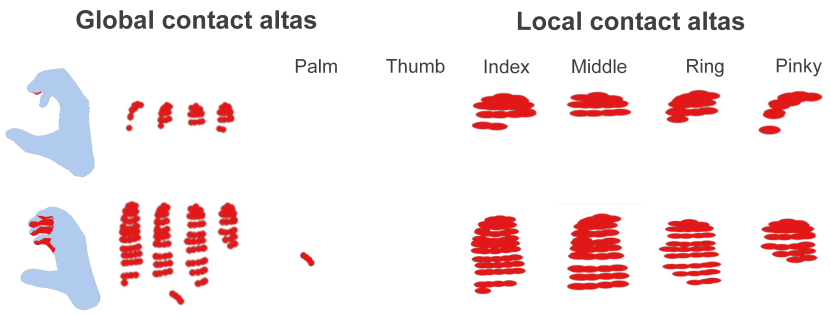


Figure 8: More contact visualization.

A.8 Statistical analysis of Dataset Distribution

We analyze regional contact patterns in the MANO vertex domain before rasterizing them into the contact atlases $\mathbf{C}$ introduced in Section 3. This avoids introducing interpolation and resolution effects into the dataset statistics.

For the i-th retained clip, let

$$\mathbf{C}_{\text{vert}}^{(i)} \in \{0,1\}^{N\times V} \tag{18}$$

denotes its binary vertex-level contact sequence, where $N = 33$ is the sequence length and $V = 778$ is the number of MANO vertices. The element $C_{\text{vert},t,v}^{(i)}$ indicates whether vertex v is in contact at frame t. The binary labels are obtained by thresholding the stored contact values at 0.5.

As in the contact atlas construction, the hand is divided into $R = 6$ palmar regions: palm, thumb, index, middle, ring, and pinky. Let $\mathcal{V}_r^{\text{palm}} \subseteq \{1,\dots,V\}$ denote the set of valid palm-side MANO vertices assigned to region r. The corresponding numbers of vertices are

$$(111,\ 52,\ 57,\ 55,\ 57,\ 62),$$

following the region order above.

Each text instruction ℓ_i is assigned to a verb category according to its first normalized word. For a verb category ν, we define

$$\mathcal{W}_\nu = \{i \mid \text{verb}(\ell_i) = \nu\}, \tag{19}$$

where $\mathcal{W}_\nu$ contains all retained clips associated with that verb.

Region Effective Contact Ratio. A hand region is considered active at frame t if at least one of its palm-side vertices is in contact. The region-wise effective contact ratio is defined as

$$\text{ECR}(\nu, r) = \frac{1}{|\mathcal{W}_\nu|N} \sum_{i\in\mathcal{W}_\nu} \sum_{t=1}^{N} \mathbf{1}\left[\sum_{v\in\mathcal{V}_r^{\text{palm}}} C_{\text{vert},t,v}^{(i)} > 0\right], \tag{20}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. $\text{ECR}(\nu, r)$ measures the fraction of frames in which region r maintains any contact and therefore characterizes temporal contact persistence independently of contact area.

Region Contact Density. Region contact density is defined over all eligible palm-side vertex frame pairs:

$$\text{RCD}(\nu, r) = \frac{\sum_{i\in\mathcal{W}_\nu} \sum_{t=1}^{N} \sum_{v\in\mathcal{V}_r^{\text{palm}}} C_{\text{vert},t,v}^{(i)}}{|\mathcal{W}_\nu|N|\mathcal{V}_r^{\text{palm}}|}. \tag{21}$$

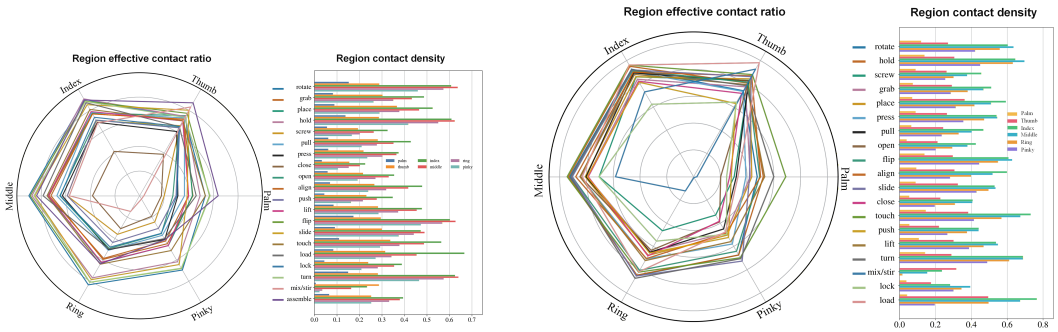
$\text{RCD}(\nu, r)$ measures the fraction of contacted vertex frame pairs within region r, jointly reflecting the spatial coverage and occurrence frequency of regional contact.

Both statistics lie in $[0, 1]$ and are computed by pooling all retained clips associated with each verb. No per-verb normalization or min-max rescaling is applied.

Visualization. The radar plot represents each verb ν by the six-dimensional vector

$$\big[\mathrm{ECR}(\nu, 1), \ldots, \mathrm{ECR}(\nu, R)\big], \tag{22}$$

with one axis for each palmar region. The grouped bar chart reports $\mathrm{RCD}(\nu, r)$ for every verb region pair. Consequently, the radar plot characterizes temporal contact persistence, whereas the bar chart characterizes regional contact occupancy.



(a) Main and ablation dataset

(b) Task-level dataset

Figure 9: **HoloAssist Verb-level regional contact statistics.** Each panel presents the region-wise effective contact ratio using a radar plot and regional contact density using a grouped bar chart. *(a)* Statistics computed over the dataset used for controlled ablation experiments using 33-frame clips with a temporal stride of 10. *(b)* Statistics computed over the task-level dataset used by LatentAct, where valid samples with large hand displacements are filtered out, using 33-frame clips with a temporal stride of 1.

A.9 Ablation Study

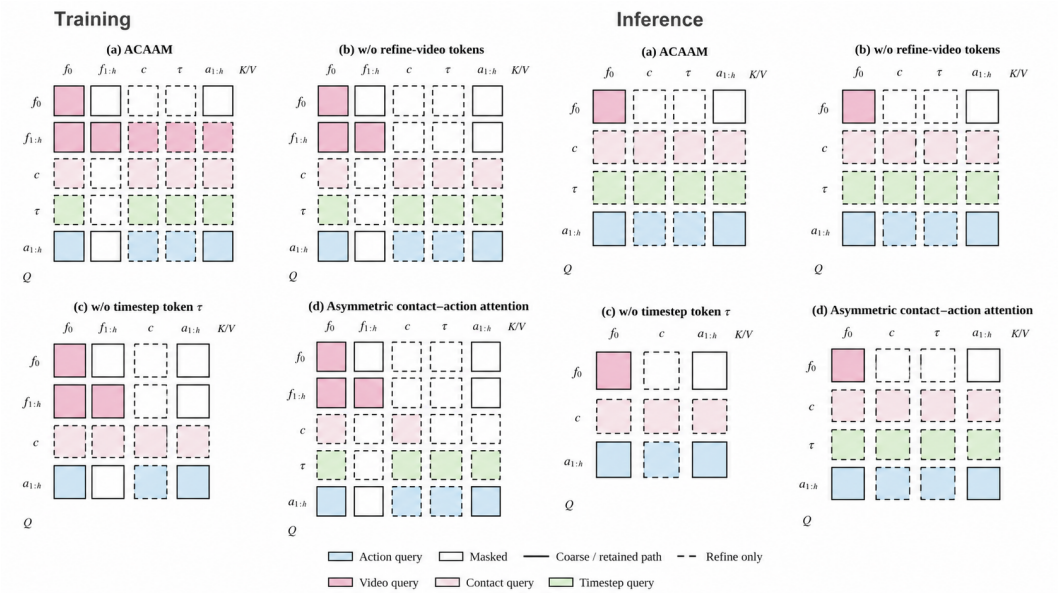


Figure 10: Visualization of ACAAM ablation studies.

ACAAM Ablations. ACAAM enforces causal dependence on visual observations while preserving bidirectional action denoising. Video tokens follow a first-frame-causal attention mask. During contact-aware refinement, action queries attend to the observed first-frame video tokens, contact skill codes, the timestep token τ, and all action tokens, but cannot access future-video keys. The

training mask is directional: future-video queries may attend to the refinement tokens for auxiliary video supervision, whereas refinement queries cannot attend to future-video tokens. At inference, future-video tokens are omitted, and action generation remains conditioned on the observed first frame and the contact-aware refinement tokens. The ACAAM ablation results for different attention masks are shown in Figure 10.

Codebook Capacity. K denotes the number of discrete entries in the image contact skill codes. The codebook with $K = 16$ remains close to WAM (Baseline) and yields no gain in the absolute trajectory metrics. As shown in Table 4, the transition from $K = 16$ to $K = 64$ reduces MPJPE, MPVE, and MWTE by 5.33%, 5.38%, and 5.65%, respectively. The trend suggests that expanded discrete codebook capacity captures fine-grained contact transition patterns. Sufficient codebook capacity preserves fine-grained regional contact transitions for action refinement.

Table 4: Ablation studies on contact-aware action refinement.

Method	K	MPJPE↓	MPVE↓	MWTE↓
WAM (Baseline)	–	8.74	8.71	8.59
Codebook capacity (image contact skill codes)				
$K = 16$	16	8.75	8.72	8.60
$K = 32$	32	8.36	8.33	8.18
$K = 64$	64	**8.29**	**8.26**	**8.11**

Training Objective. As shown in Table 5, the loss term ablation is removed from CAD-WAM with codebook size $K = 64$. Removing the reference-space loss $\mathcal{L}_{\text{ref}}$ increases MPJPE, MPVE, and MWTE by 0.39, 0.39, and 0.40 cm, respectively. Removing the refinement-stage auxiliary video loss $\mathcal{L}_{\text{vid}}^{\text{aux}}$ increases the corresponding errors by 0.18, 0.20, and 0.30 cm. These results support the contributions of reference-space supervision and refinement-stage video supervision to hand trajectory accuracy. In addition, both video losses use the same ground-truth videos but supervise separate forward passes. The base loss $\mathcal{L}_{\text{FM}}^{\text{vid}}$ remains active in both configurations, whereas $\mathcal{L}_{\text{vid}}^{\text{aux}}$ additionally propagates gradients through ACAAM to the contact-aware refinement expert. The auxiliary objective therefore differs from a simple increase in the base video-loss weight.

Table 5: Ablation study of the training objective on HoloAssist with codebook size $K = 64$.

Method	MPJPE↓	MPVE↓	MWTE↓
CAD-WAM	8.10	8.06	7.81
w/o $\mathcal{L}_{\text{ref}}$	8.49	8.45	8.21
w/o $\mathcal{L}_{\text{vid}}^{\text{aux}}$	8.28	8.26	8.11

As shown in Table 6, each training sample consists of 33 frames and uses the same filtering criteria as LatentAct, excluding zero-contact samples and valid samples with large hand displacements. CAD-WAM outperforms the existing state-of-the-art methods. The training dataset excludes zero-contact samples and valid samples with large hand displacements.

Table 6: Quantitative comparison. LatentAct* denotes the results obtained under the same dataset split with $T = 30$. CAD-WAM outperforms the other methods.

Dataset	Method	MPJPE↓	PA-MPJPE↓	MPVE↓	PA-MPVE↓	MWTE↓	MRE↓
HoloAssist	LatentAct* (Prakash et al., 2025)	7.61	2.99	-	-	-	-
	LatentAct (Prakash et al., 2025)	8.15	1.00	8.05	0.94	7.13	0.328
	– no concat map	7.99	1.03	7.89	0.98	6.86	0.338
	LatentAct-Diff (Prakash et al., 2025)	8.41	1.14	8.33	1.09	7.52	0.363
	– no concat map	6.93	1.12	6.83	1.07	6.00	0.361
	MEgoHand (Zhou et al., 2025)	5.70	1.11	5.56	1.06	3.75	0.354
	CAD-WAM	**3.62**	**0.60**	**3.57**	**0.57**	**3.12**	**0.171**
ARCTIC	LatentAct*(Prakash et al., 2025)	14.77	3.76	-	-	-	-
	LatentAct (Prakash et al., 2025)	12.29	1.72	12.22	1.66	11.42	0.797
	– no contact map	11.97	1.40	11.93	1.37	11.38	0.628
	LatentAct-Diff (Prakash et al., 2025)	16.95	1.59	16.91	1.55	16.02	0.730
	– no contact map	14.32	1.29	14.23	1.25	12.93	0.622
	MEgoHand (Zhou et al., 2025)	10.70	1.54	10.66	1.51	10.04	0.547
	CAD-WAM	**4.75**	**0.76**	**4.71**	**0.75**	**4.22**	**0.272**