CONTACTFLOW: A VIDEO ACTION CONDITIONING THAT TRANSFERS ACROSS EMBODIMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

World models offer a promising route toward more general embodied intelligence by enabling robotic agents to anticipate the consequences of actions before execution. Video Generation based world models are particularly attractive because they can learn rich spatiotemporal interaction dynamics from large-scale video data. By capturing aspects of real-world physics, these dynamics could allow the models to simulate how scenes evolve under different actions. Realizing this potential requires models to learn from interaction data at sufficient scale and diversity, most of which depict humans, and transfer the acquired knowledge to different robotic embodiments. However, such a transfer requires an action representation that provides a common interface across humans and robots. Existing approaches instead condition on embodiment-specific actions, poses, or kinematics, tying the learned dynamics to a particular actor or control space. This coupling limits both the use of heterogeneous interaction data and transfer across embodiments. We propose *Contact Flow*, an embodiment-agnostic action representation that encodes manipulation through the spatiotemporal trajectories of 3D contact points between an actor and a target object. Contact Flow captures the object-side interaction geometry that determines how actions affect the environment, providing a common interface across embodiments. We condition a video world model on Contact Flow to predict future scene evolution from an initial observation. Experiments across human and robotic manipulation benchmarks show that Contact Flow enables accurate video prediction and effective cross-embodiment transfer.

1 INTRODUCTION

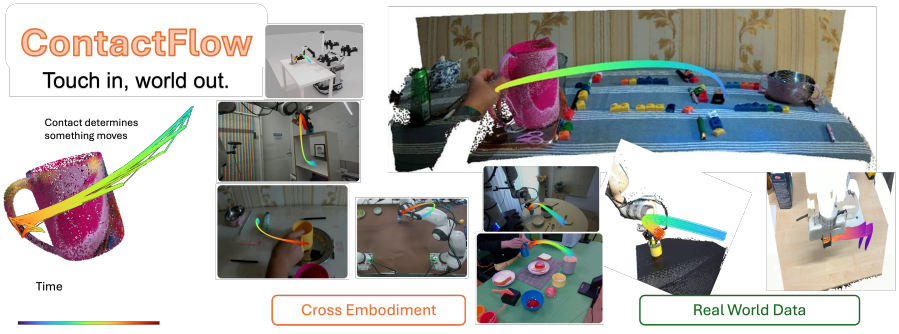


Figure 1: **Contact Flow overview.** Contact Flow encodes physical manipulation through the spatiotemporal trajectory of 3D contact points between an actor and environments.

Robots that are asked to fulfill open-ended tasks need to anticipate how their actions will affect the environment around them. Video world models provide this capability by predicting future video frames from past observations and candidate actions, enabling agents to verify their planned actions and compare different options before execution.

However, video world models require a large amount of video data to learn the effects of different actions: Their training data does not only need to cover the visual diversity of potential deployment

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.

environments, but also the space of all possible actions and interactions with many different objects and parts of these environments. Therefore, a video world model can achieve higher accuracy if it can be trained on videos collected with many different embodiments. Ideally, this includes the large scale of available videos of humans interacting with the world.

The observable dynamics from interactive videos of any agent can be divided into two categories: Active and passive dynamics. We refer to active dynamics as the motion of the agent in the environment. In general, any embodiment is restricted in its movement by its kinematic model, the overall size of the agent, its material, strength, etc. Active dynamics are therefore specific to the acting agent. With passive dynamics instead we refer to the environment response to an interaction, e.g. an object moving when pushed or a door opening when pulled. Passive dynamics are independent of the agent and instead governed by physical laws such as mechanical constraints or material properties.

In this work, we aim to decouple active and passive dynamics through a generalizing action encoding. We argue that a world model is most useful if its training focuses on the passive dynamics, i.e. learning the physical laws that govern the outcome of an interaction. Hence, the input and output should be as independent of the acting agent as possible. Such independence has multiple advantages: As argued above, cross-embodiment training taps into a larger scale of possible training data, pushing the generalization of the learned model. Furthermore, it allows the same model to be re-used in deployment on different robots.

The interface between active and passive dynamics is the contact of the agent with the environment. The resulting object motion depends on where contact occurs and how the contact areas move over time. Contact geometry and contact motion are visually accessible and capture the part of the interaction that is shared across embodiments. This makes object-side contact a natural interface for learning passive dynamics. We formalize this intuition as **Contact Flow**, the trajectory of 3D contact points between actor and object projected into image space (Fig. 1). Contact Flow discards everything about the actor except the locus of interaction. It transfers across embodiments because it is expressed entirely in terms of the object surface.

In this work, we train a video world model conditioned on Contact Flow extracted from both human and robot interaction videos. We then demonstrate video prediction on held-out human and robotic datasets, including previously unseen robot embodiments, and we use the model to evaluate action plans on a real-world tabletop robot. Our results demonstrate that contact flow achieves better generalization over different embodiments than other action conditionings. In summary, our contributions are:

- **Contact Flow**, an object-side conditioning representation that encodes interactions through 3D contact-point trajectories to support transfer across embodiments.
- A curated multi-embodiment dataset (**ContactFlow-212K**) with a data-processing pipeline for extracting Contact Flow from heterogeneous human and robot demonstrations and candidate robot trajectories.
- A **Contact Flow-conditioned video world model** with three mechanisms for injecting sparse contact trajectories into a shared video backbone.
- A multi-dataset evaluation of cross-embodiment video prediction on held-out human and robot datasets, together with a qualitative real-world action-verification deployment.

2 Related Work

Generative video models for robot learning. Generative video models have been used as visual planners, policy learners, data generators, and interactive simulators for robotics. UniPi predicts text-conditioned future videos as visual plans and recovers executable actions with an inverse-dynamics model Du et al. (2023). GR-1 instead jointly predicts future images and robot actions after large-scale video pretraining Wu et al. (2024), while RoboDreamer factorizes language into action and spatial primitives to improve compositional generalization of imagined robot rollouts Zhou et al. (2024). Other works attempt to learn policies from demonstration. Dreamitate transfers human tool-use videos to a robot through the tracked 6-DoF trajectory of a shared rigid tool Liang et al. (2025), DreamGen instead adapts a video model to a target robot embodiment and pseudo-labels generated videos for policy training Jang et al. (2025). Other systems use video models to roll out

explicit robot actions. IRASim predicts long-horizon observations from action trajectories Zhu et al. (2025), Ctrl-World supports policy-in-the-loop simulation and synthetic policy improvement Guo et al. (2025), and Interactive World Simulator uses autoregressive action-conditioned rollouts for demonstration generation, policy training, and evaluation Wang et al. (2026). STORM searches over VLA-proposed action sequences by evaluating their predicted visual futures Lin et al. (2025), whereas DRAW2ACT generates demonstrations from depth-aware robot and object trajectories Bai et al. (2025). Collectively, these works connect video prediction to action, but condition on semantic goals, target-embodiment actions, or actor and tool trajectories rather than the local contact interface.

Control interfaces for interaction video generation. The conditioning representation determines which aspects of an interaction a video model can control and transfer. At a high level, language specifies goals or intermediate plans in Video Language Planning, DreamGen, and TesserAct Du et al. (2024); Jang et al. (2025); Zhen et al. (2025). Embodiment specific models instead consume frame-aligned action trajectories, gripper-coordinate commands or body poses Zhu et al. (2025); Agarwal et al. (2025); Guo et al. (2025); Pallotta et al. (2026). Visual controls provide strong inductive bias for training but expose spatial structure more directly. InterDyn supplies the mask trajectory of a driving entity and generates its effects on the scene Akkerman et al. (2025); Mask2IV predicts actor and object mask trajectories from target and action cues before using them to synthesize an interaction Li et al. (2026). TASTE-Rob refines hand-pose trajectories to generate task-oriented human demonstrations Zhao et al. (2025), while Visual Action Prompts and OSCAR render human hands or robot kinematics as skeleton trajectories, providing a shared but still actor-centric interface across embodiments Wang et al. (2025d); Wu & Gao (2026). More generic video-control methods use camera flow, reference-video optical flow, or arbitrary point tracks Jin et al. (2025); Koroglu et al. (2025); Geng et al. (2025). All these discussed signals are either too high-level, embodiment-specific, or tied to the actor's visible morphology. In contrast, Contact Flow provides a contact geometry between actor and object, providing a local and embodiment-agnostic representation of the interaction.

Hand-object contact representations. Contact has long provided a compact description of how hands and robots interact with objects. ContactDB recovers object-surface grasp contact through thermal imaging Brahmbhatt et al. (2019), and ContactPose pairs detailed contact maps with 3D hand and object poses Brahmbhatt et al. (2020). ContactOpt predicts desirable contact and optimizes a reconstructed hand pose to realize it Grady et al. (2021). H2O complements these resources with synchronized two-hand and object poses and interaction labels, but does not annotate contact Kwon et al. (2021). In robotics, Contact-GraspNet parameterizes 6-DoF grasp proposals by candidate scene contacts Sundermeyer et al. (2021), while quasi-dynamic contact models support global planning for contact-rich manipulation Pang et al. (2023). Thus, prior work primarily treats contact as measured supervision, an estimation or optimization target, a grasp parameterization, or a dynamics model. We instead use estimated contact geometry as a control input to a generative video model.

Cross-embodiment interaction representations. Prior cross-embodiment interfaces can be organized by whether they retain the actor or abstract it away. Kinema4D projects a URDF-driven robot trajectory into spatiotemporal pointmaps Xu et al. (2026), and BridgeV2W renders robot actions as embodiment-mask sequences Chen et al. (2026). Both expose a common image-aligned interface while retaining the acting robot's geometry. Visual Action Prompts and OSCAR similarly encode human or robot kinematics as rendered skeletons Wang et al. (2025d); Wu & Gao (2026). In contrast, Im2Flow2Act removes the actor and predicts manipulated-object flow Xu et al. (2024), 3DFlowAction predicts object-centric 3D flow and optimizes robot actions to realize it Zhi et al. (2025). Object-centric 3D Motion Field transfers dense object motion estimated from human videos to robot control Yin et al. (2026). Any-point Trajectory Modeling predicts tracks for arbitrary selected image points and is therefore trajectory-centric rather than inherently object-centric Wen et al. (2024). These representations encode actor motion, tool motion, arbitrary tracks, or the resulting motion of an object. Contact Flow instead conditions generation on the localized actor-object interface that induces that motion, abstracting interaction dynamics from the visible morphology of a particular actor.

3 CONTACT FLOW

We first formalize Contact Flow as an action representation and describe how we condition and train our video world model on it. We then present the data-processing pipeline used to extract Contact Flow from existing human and robotic demonstration datasets for model training (Fig. 2).

Contact Flow represents an interaction through trajectories of 3D contact points between an actor and the environment (Fig. 4). At each time step t, we represent Contact Flow as a set of contact points

$$\mathbf{C}_t = \left\{ \mathbf{c}_t^{(i)} \right\}_{i=1}^{N_t}, \qquad \mathbf{c}_t^{(i)} = (x, y, z, \Delta x, \Delta y, \Delta z, w) \in \mathbb{R}^7. \tag{1}$$

Here, (x, y, z) denotes the 3D position of contact point i at time t, $(\Delta x, \Delta y, \Delta z)$ denotes its displacement to the following time step, and $w \in [0, 1]$ is the confidence assigned to the estimated contact and displacement.

Contact Flow retains the object-side interaction geometry relevant to the environment response while omitting actor-specific motion. For example, a person turns a door handle through the contacts established with the handle and the motion of those contacts over time. The same object-side interaction can therefore be specified without encoding the actor's body or control space. Different hands or grippers may realize an action through different contact trajectories, but all trajectories use the same representation. We hypothesize that sufficiently diverse training data exposes the world model to these variations and enables it to generalize to previously unseen Contact Flows and embodiments.

3.1 CONTACT FLOW VIDEO GENERATION WORLD MODEL

We instantiate the Contact Flow Video Generation World Model as a latent video Diffusion Transformer (DiT) based on Wan Wang et al. (2025a). Given an initial frame x_0 and a Contact Flow sequence $\mathbf{C}_{1:T}$, the model predicts the future video $\mathbf{x}_{1:T}$. A frozen video VAE encoder $\mathcal{E}$ maps $\mathbf{x}_{0:T}$ to latent representations $\mathbf{z}_{0:T} = \mathcal{E}(\mathbf{x}_{0:T})$, and the world model learns the conditional distribution $p_\theta(\mathbf{z}_{1:T} \mid z_0, \mathbf{C}_{1:T})$, where z_0 represents the initial observation and $\mathbf{C}_{1:T}$ specifies the interaction over the prediction horizon. We train the model with the flow-matching objective of the Wan backbone. During training, z_0 remains clean while noise is applied to the future latents $\mathbf{z}_{1:T}$. The velocity network is conditioned on both z_0 and $\mathbf{C}_{1:T}$ and learns the velocity field of the future latent sequence. At test time, we initialize the future latents with noise and integrate the learned velocity field conditioned on the initial frame and the requested Contact Flow. The VAE decoder then maps the resulting latents $\hat{\mathbf{z}}_{1:T}$ to predicted frames $\hat{\mathbf{x}}_{1:T} = \mathcal{D}(\hat{\mathbf{z}}_{1:T})$. To provide Contact Flow as an image-aligned control signal, we project each 3D contact point $\mathbf{c}_t^{(i)}$ into the image using the camera intrinsics $\mathbf{K}$ as $\mathbf{u}_t^{(i)} = \pi(\mathbf{K}, (x, y, z))$. We consider three mechanisms for injecting the projected Contact Flow while keeping the remaining model configuration fixed to isolate the effect of the conditioning mechanism. All variants receive the clean first-frame latent together with a log-depth latent, its validity channel, and a per-slot Contact Flow state, and use low-rank adapters in the attention and feed-forward layers. The state encodes the sampling interval, time since release, contact density $\log(1 + n)$ with mean per-frame point count n, and the contact phase, which is pre-contact, touching, or released. A zero-initialized linear layer adds the state to the timestep embedding. During training, we weight the object region by 4, downweight regions around the actor, and assign zero loss to latent slots before the first contact.

Contact tokens. We pool contact points within each spatiotemporal latent cell and encode every occupied cell as a token containing its mean 3D position, mean displacement, confidence, subcell offset, and depth features. These tokens share the cell's 3D rotary position encoding, participate in self-attention during the first ten transformer blocks, and are then removed.

Contact pointmap video. We rasterize each frame's contact positions and displacements as image-space discs under a signed-log encoding, then map the resulting pointmap videos to latent features with the frozen VAE. We concatenate these features and a pooled occupancy map with the noisy video latent and use a zero-initialized widened patch embedding to preserve the pretrained backbone at initialization.

Contact latent transport. Following ATI Wang et al. (2025b) and Wan-Move Chu et al. (2025), we sample the clean first-frame latent at each initial contact position and transport this appearance feature along the corresponding Contact Flow trajectory. At later time steps, we splat the transported

features onto the latent grid and hold them at the final contact position after release, so the control uses no visual information from future frames.

3.2 CONTACTFLOW-212K DATASET.

We construct **ContactFlow-**$212K$, a dataset of manipulation sequences with Contact Flow annotations derived from heterogeneous human, real-robot, and simulated robot interactions, as shown in Fig. 2. The source datasets in each category are listed in Appendix B. Together, they provide a wide variation in embodiments, viewpoints, and interaction geometry. To obtain a common representation from these heterogeneous sources, we process each sequence with the same Contact Flow extraction framework. We use dataset-provided annotations and modalities whenever available and estimate only missing quantities, allowing datasets with different annotation sets and embodiments to contribute to the same training corpus. We retain only sequences recorded with stationary cameras so that image-space motion reflects scene and actor motion rather than camera motion.

Contact Flow extraction. We first recover scene geometry by estimating depth with Foundation-Stereo Wen et al. (2025) and unprojecting it into camera-frame pointmaps using the calibrated camera intrinsics. We combine these pointmaps with available object poses, robot poses, segmentation masks, and camera annotations to reconstruct the manipulated object, estimate actor-object contact, and recover temporally consistent contact trajectories, as summarized in Fig. 3. Appendix B describes the available annotations, camera calibration, and dataset-specific preprocessing in detail. **Object reconstruction.** Although the pointmap captures the scene geometry, unrelated depth fragments can create spurious contact detections. We therefore isolate the manipulated objects before estimating contact. To identify the object, we query Gemini Team et al. (2023) using the available task text and video. We then use the predicted object names to prompt SAM3 Carion et al. (2025), which segments and tracks the object across frames. Finally, we lift each 2D object mask into 3D using the corresponding pointmap. This yields the object points used for contact estimation. Appendix B describes the dataset-specific inputs, camera calibration, and depth processing. **Contact Area estimation.** Using the tracked object masks from the object-model stage, we lift the object geometry into 3D in the initial and final frames of the interaction. We compute candidate contacts on both geometries using the embodiment-specific hand or gripper estimates detailed in Appendix B. We then align the initial and final object geometries with iterative closest point (ICP) under a local-rigidity assumption and transfer the contacts through this alignment. This local alignment also accommodates non-rigid objects when deformation near the contacted region is limited. **Contact Flow estimation.** Recovering Contact Flow requires associating each contact point with the same physical point at subsequent time steps. These associations are ambiguous when individual surface points lack distinctive visual features. For example, a uniformly colored object may provide reliable point tracks only along its boundary because its interior points have similar appearances. To estimate these correspondences, we apply a Hough-based filter over an m-time-step window. The filter favors nearby points with consistent motion and rejects isolated or inconsistent matches. We assign each retained correspondence a confidence score w that combines its local spatial and temporal consistency with an embodiment-specific contact estimate. For robot data, this estimate is derived from the distance between the gripper and the object surface. For human data, it combines the HaMeR hand-pose and HACO contact confidences.



Figure 2: **ContactFlow-212K dataset composition.** Human, real-robot, and simulated interaction data used to construct our multi-embodiment training corpus.

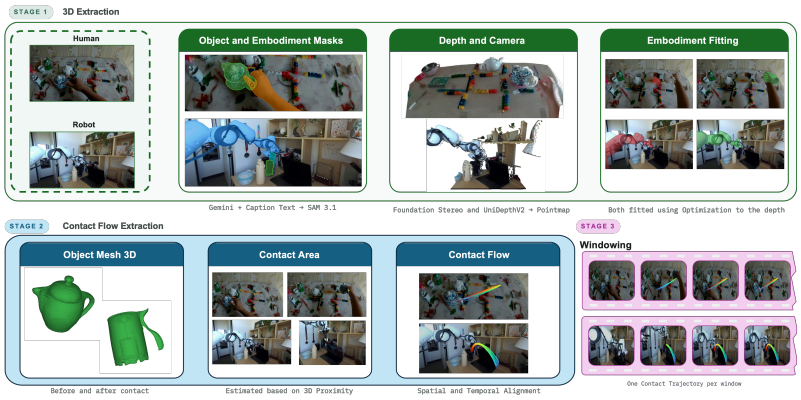
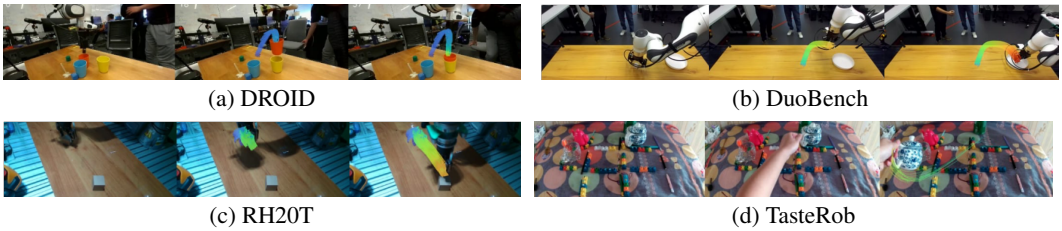


Figure 3: **Data preprocessing pipeline.**



(a) DROID (b) DuoBench

(c) RH20T (d) TasteRob

Figure 4: **Contact Flow trajectory examples.** Contact-point trajectories projected into image space for DROID (a), DuoBench (b), RH20T (c), and TasteRob (d).

4 EXPERIMENTS

To evaluate Contact Flow as an embodiment-independent action-conditioning signal for video generation, we ask three questions: (1) Does Contact Flow enable a video model to predict the visual consequences of an action? (2) Can it represent actions across human and robot embodiments? (3) Does joint training across embodiments improve generalization to unseen domains and embodiments? We test the first two questions by training Contact Flow-conditioned models on diverse human and robot manipulation data and evaluating them on held-out, out-of-distribution datasets against published baselines. To isolate the effect of cross-embodiment training, we compare models trained on human data, robot data, and their mixture. We also compare Contact Flow with alternative conditioning representations under matched training data and backbone settings.

Evaluation protocol. To isolate the passive response of the environment from the visual appearance of the acting embodiment Chang et al. (2023), we mask the visible robot arm or human hand in both the prediction and ground truth using SAM 3.1 Carion et al. (2025). We measure reconstruction fidelity with PSNR, perceptual similarity with LPIPS Zhang et al. (2018), and higher-level visual similarity with DreamSim Fu et al. (2023). We evaluate target-object motion using object-mask IoU averaged over the trajectory, End IoU averaged over the final evaluated frames, and final object-position error normalized by the initial object bounding-box diagonal. We additionally report the missing-object rate as a detection-based failure measure related to that used for image generation Qiu et al. (2022). For all trained models, we use a Wan2.2-5B backbone and train for 5k steps with a batch size of 32.

Datasets. We train on sampled subsets of the three data sources shown in Fig. 2, comprising human demonstrations, simulated robot interactions, and real-robot data, while excluding all samples with camera motion. Additional details on the subset composition are provided in the appendix. We use separate datasets exclusively for evaluation, with no overlap with the training data. Specifically, we evaluate out-of-distribution human interactions on 100 HO-Cap Wang et al. (2025c) samples and 50 HOGraspNet samples, and out-of-distribution robot interactions on 100 AgiBot Bu et al. (2025) samples and 100 DuoBench real Jülg et al. (2026) samples. We evaluate in-domain robot interactions on 100 DROID Khazatsky et al. (2024) samples, 100 RH20T validation Fang et al. (2024) samples, and 50 RoboMIND validation samples.Each sample corresponds to a complete contact trajectory, which we temporally resample to 49 frames and spatially resize to 832×480 pixels.

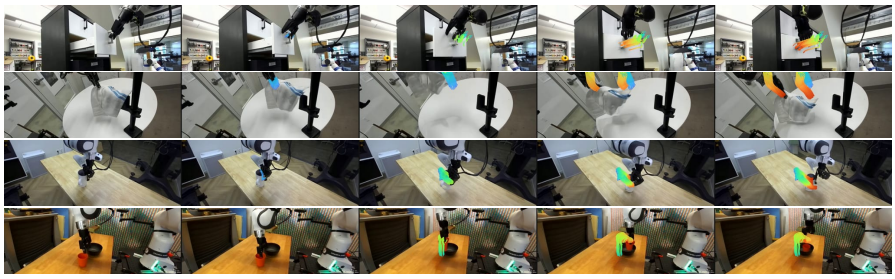


Figure 5: **Generated rollouts.** The Contact Flow extracted from each real interaction is provided as input, and the generated object motion follows its direction and spatial trajectory.

Baselines. We compare Contact Flow against six baselines spanning different conditioning paradigms. TesserAct Zhen et al. (2025) predicts future observations from an initial image and task description, testing whether a goal specification is sufficient without an explicit action representation. CTRL-World Guo et al. (2025) uses frame-level robot actions and was trained on DROID, providing a robot-specific comparison. OSCAR Wu & Gao (2026) and Kinema4D Xu et al. (2026) encode the acting embodiment as 2D kinematic skeletons and 4D pointmaps, respectively, testing whether cross-embodiment actor geometry captures the interaction. ATI Wang et al. (2025b) injects user-specified point trajectories into the video model, whereas Wan-Move Chu et al. (2025) transports first-frame latent features along dense point trajectories. These methods test generic trajectory control without an explicit representation of contact. To disentangle the effect of these conditioning representations from differences in backbone architecture and training data, we additionally implement controlled OSCAR-like and Kinema4D-like variants with the same backbone and training data as Contact Flow, conditioning them on 2D kinematic skeletons and 4D embodiment pointmaps, respectively.

4.1 Contact Flow for Video Generation Across Embodiments and Domains

We first examine whether Contact Flow provides a common action-conditioning interface across different interaction types. Figure 5 shows qualitative rollouts from the 5B model trained on human and simulated robot interactions. Across pushing, pulling, sliding, and pick-and-place interactions, the generated object motion clearly follows the provided Contact Flow in its direction and spatial trajectory. This correspondence indicates that Contact Flow provides effective spatial control over diverse interactions without explicitly encoding the acting embodiment. We next examine whether Contact Flow can represent actions across different embodiments. We evaluate our models zero-shot on unseen human and real-robot domains and compare against strong video prediction baselines, including models with larger backbones and substantially longer training. As shown in Table 1, our models remain competitive across controllability, visual fidelity, and semantic fidelity, and consistently outperform the published baselines across both evaluation domains. Importantly, this generalization extends across fundamentally different embodiments. The same Contact Flow representation supports prediction for unseen human interactions as well as for previously unseen robot embodiments without embodiment-specific adaptation. These results show that Contact Flow provides a shared action representation that transfers across human and robot embodiments and enables a single video world model to predict diverse manipulation interactions beyond the embodiments observed during training. We further compare different mechanisms for incorporating Contact Flow into the video world model. While no single encoding performs best across all evaluation metrics, contact tokens provide the most consistent control over the final interaction outcome across both unseen human and robot embodiments. Pointmap conditioning yields stronger spatial object overlap, while latent transport favors visual fidelity. We therefore adopt contact tokens as our primary variant, as accurately reaching the prescribed interaction outcome is most directly aligned with our goal of controllable action-conditioned prediction. Overall, the results suggest that different Contact Flow encodings emphasize complementary aspects of generation, while contact tokens provide the strongest balance for interaction control across embodiments.

Application to Action Outcome Prediction As a real-world application, we use $\pi_{0.5}$ Physical Intelligence et al. (2025), which achieves 100% task success in simulation but 0% on the real robot. We convert ten simulated trajectories into Contact Flow and generate ten real-world rollouts per trajectory from the first real camera frame. We then execute the same ten trajectories on the real

Table 1: **Comparison with published video world models.** Contact Flow consistently outperforms published baselines on unseen human and robot embodiments.

Method	Human					Robot				
	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓
Contact tokens (ours)	0.431	0.484	**0.148**	22.14	0.057	0.325	0.243	**0.438**	25.75	0.046
Contact pointmap video (ours)	**0.466**	**0.532**	0.149	21.74	0.058	**0.332**	0.246	0.490	25.63	**0.044**
Contact latent transport (ours)	0.429	0.464	0.169	**22.19**	**0.055**	0.326	**0.252**	0.532	**26.07**	**0.044**
TesserAct Zhen et al. (2025)	0.330	0.339	0.525	17.36	0.120	0.163	0.029	1.568	12.88	0.170
CTRL-World Guo et al. (2025)	0.308	0.312	0.670	17.96	0.234	0.130	0.027	1.674	21.73	0.154
OSCAR Wu & Gao (2026)	0.335	0.328	0.468	21.34	0.099	0.233	0.119	0.960	22.59	0.097
Kinema4D Xu et al. (2026)	0.320	0.287	0.588	19.81	0.093	0.240	0.129	0.910	16.42	0.097
ATI Wang et al. (2025b)	0.420	0.388	0.297	14.69	0.200	0.264	0.132	0.684	16.01	0.244
Wan-Move Chu et al. (2025)	0.418	0.386	0.245	15.51	0.148	0.254	0.142	0.808	20.14	0.095

Table 2: **Contact flow conditioning across training data.** Contact Flow transfers from human and simulated interactions to unseen robots more effectively than actor-centric conditioning.

Training data	Conditioning	Human					Robot				
		IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓
Human	**Contact tokens (ours)**	0.431	0.490	0.158	22.08	**0.057**	0.241	**0.163**	**0.732**	25.30	0.054
	Contact pointmap video (ours)	0.464	**0.541**	**0.154**	21.18	0.061	**0.260**	0.162	0.749	23.37	0.059
	Contact latent transport (ours)	0.409	0.468	0.175	21.80	0.061	0.234	0.131	0.808	**25.57**	**0.051**
	Actor pointmap baseline	**0.477**	0.449	0.314	**22.49**	0.064	0.137	0.033	1.427	23.74	0.066
	Actor skeleton baseline	0.442	0.432	0.334	19.88	0.075	0.140	0.045	1.497	16.35	0.155
Human + Sim	**Contact tokens (ours)**	0.438	0.499	0.160	**21.97**	**0.059**	0.314	0.230	0.478	**26.11**	**0.044**
	Contact pointmap video (ours)	0.477	**0.509**	**0.158**	21.86	0.062	**0.337**	**0.250**	**0.445**	25.62	0.046
	Contact latent transport (ours)	0.409	0.463	0.188	21.79	0.061	0.301	0.201	0.566	25.69	0.050
	Actor pointmap baseline	**0.494**	0.441	0.369	21.80	0.067	0.125	0.044	1.531	21.58	0.101
	Actor skeleton baseline	0.442	0.405	0.384	21.96	0.063	0.199	0.120	1.061	23.46	0.060
Human + Sim + Robot	**Contact tokens (ours)**	0.431	0.484	**0.148**	22.14	0.057	0.325	0.243	**0.438**	25.75	0.046
	Contact pointmap video (ours)	0.466	**0.532**	0.149	21.74	0.058	**0.332**	0.246	0.490	25.63	**0.044**
	Contact latent transport (ours)	0.429	0.464	0.169	22.19	0.055	0.326	**0.252**	0.532	**26.07**	**0.044**
	Actor pointmap baseline	**0.491**	0.458	0.286	**23.02**	**0.052**	0.169	0.072	1.315	23.93	0.060
	Actor skeleton baseline	0.438	0.416	0.331	21.38	0.064	0.250	0.158	0.869	21.09	0.091

robot and compare the generated outcomes with the observed executions. Across 100 generations, our model correctly forecasts the real-world outcome for all 70 samples from successful trajectories and 20/30 samples from failed trajectories.

4.2 Does ContactFlow improve the cross embodiment generalization of video generation

The previous comparison involves different backbones and training setups. We therefore isolate the action representation by training all conditioning approaches under the same setting in Table 2. Within the human domain, Contact Flow performs on par with actor-centric conditioning across perceptual metrics, demonstrating strong generalization to unseen human interactions despite representing only the interaction rather than the full actor geometry. Actor-centric representations similarly generalize to unseen interactions within the human embodiment, but transfer substantially less effectively when the embodiment changes. Contact Flow, in contrast, transfers directly from human interactions to unseen robots. Notably, this human-to-robot transfer is stronger than the robot-to-robot transfer of the actor-centric baselines. Adding simulated interactions further brings performance close to training with real-robot data. These results show that Contact Flow preserves strong within-embodiment generalization while extending it across embodiments, allowing a single video world model to learn from scalable human and simulated interactions data and generalize to previously unseen robots.

Ablation analysis. We study the main design choices of Contact Flow in Table 3. Removing the down-weighting of the actor has little effect on held-out interactions from the training domains but weakens prediction of object motion on unseen robots, while removing the increased object weighting leads to smaller and less consistent changes. We next examine whether additional geometric information improves interaction prediction. Adding a depth prediction head consistently reduces controllability and visual quality, whereas providing the initial object mask yields modest improvements in object localization and perceptual quality. Removing the initial depth has only a small effect, suggesting that Contact Flow does not strongly depend on accurate depth initialization. We

Table 3: **Contact-token ablations.** Combining Contact Flow with actor-centric inputs improves in-domain performance but reduces generalization to unseen embodiments.

Method	In domain					Out of domain									
	Robot (DROID, RH20T val, RoboMIND val)					Human (HO-Cap, HOGraspNet)					Real robot (DuoBench real, AgiBot)				
	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓	IoU↑	End IoU↑	Final-pos. err.↓	PSNR↑	DreamSim↓
Contact tokens (default)	0.373	0.259	0.474	24.53	0.062	0.431	0.484	**0.148**	22.14	0.057	0.325	0.243	0.438	25.75	0.046
Actor input															
+ actor skeleton	**0.436**	**0.322**	**0.459**	23.19	0.075	0.508	0.534	0.155	22.47	0.057	0.319	0.213	0.624	23.60	0.059
+ actor pointmap	0.404	0.280	0.487	24.22	0.062	**0.537**	**0.549**	0.157	**22.83**	**0.054**	0.285	0.168	0.819	25.19	0.046
Loss															
w/o actor down-weighting	0.379	0.256	0.514	24.49	0.062	0.422	0.489	0.169	22.01	0.060	0.314	0.238	0.482	25.75	0.047
w/o object up-weighting	0.369	0.243	0.542	24.68	**0.060**	0.415	0.466	0.191	21.97	0.057	0.321	**0.251**	**0.434**	25.75	0.050
+ pre-contact loss	0.373	0.246	0.499	24.59	0.061	0.437	0.495	0.158	22.15	0.056	0.312	0.216	0.491	25.67	0.051
+ depth head	0.363	0.247	0.545	23.36	0.078	0.415	0.456	0.193	21.51	0.073	0.301	0.221	0.605	24.40	0.062
Inputs															
+ frame-0 object mask	0.383	0.257	0.543	24.66	0.062	0.452	0.505	0.157	22.23	0.055	**0.327**	0.243	0.483	**26.06**	**0.045**
w/o frame-0 depth	0.381	0.253	0.513	**24.75**	0.061	0.443	0.503	0.163	22.16	0.058	0.316	0.236	0.497	25.81	0.048
w/o slot state	0.374	0.258	0.525	24.42	0.063	0.426	0.482	0.184	22.07	0.058	0.316	0.243	0.496	25.78	0.050

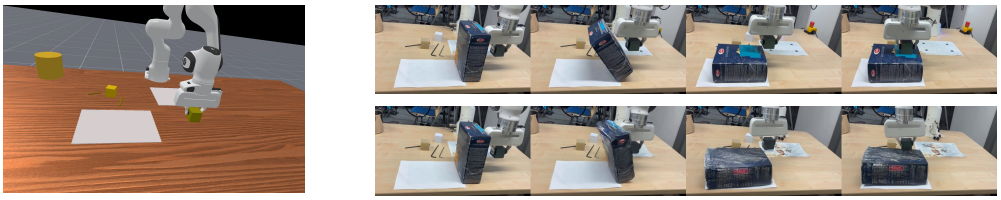


Figure 6: Real Robot Trajectory verification.

also find no consistent improvement from removing the slot state or adding a pre-contact objective. Finally, we examine whether explicit actor geometry complements Contact Flow. Adding actor skeletons or pointmaps improves alignment with the observed interaction in-domain and on unseen human interactions, but substantially reduces the accuracy of the predicted final object state on unseen robots. This indicates that actor-centric signals provide strong cues when the embodiment remains similar, while Contact Flow better preserves the interaction dynamics required for transfer across embodiments. Overall, our ablations support the default formulation and show that suppressing embodiment-specific information is particularly important for cross-embodiment prediction.

Error Analysis. The main failures arise from rendering and physical reasoning. Grippers are sometimes rendered with blurred geometry, and human predictions can introduce spurious hand artifacts that affect nearby object regions. Errors increase for small objects and interactions that require precise spatial constraints. For example, an object may follow the prescribed contact motion toward a cup but appear behind it instead of inside it. We also observe some failures in basic dynamics, such as unsupported objects remaining suspended instead of falling under gravity. Contact Flow therefore provides strong control over interaction motion, while fine geometry and physics remain hard.

5 Conclusion and Limitations

Contact Flow provides a shared action representation for video prediction across human and robot embodiments by encoding manipulation at the actor-object contact interface. Across the evaluated interactions, Contact Flow preserves the intended object dynamics under changes in appearance and kinematics and transfers more reliably across embodiments than actor-centric conditioning. Contact tokens provide the strongest control over the final object state, while pointmap and latent conditioning offer complementary strengths in spatial alignment and visual fidelity. Our sim-to-real application further shows how Contact Flow can connect actions generated in simulation to predictions of their physical consequences before execution. The current formulation relies on the video backbone to translate the explicit Contact Flow geometry into spatially and physically consistent dynamics, while scene reconstruction, contact estimation, and inference speed remain important components of the overall pipeline. Stronger video backbones and faster inference provide natural opportunities to extend this capability to more complex interactions and longer horizons. Overall, Contact Flow enables video world models to learn shared interaction dynamics from scalable human and simulated data and transfer them across embodiments.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and discover related work, support research execution through agentic coding, and generate synthetic data using vision-language models as part of our data pipelines. We reviewed and verified all AI-assisted outputs and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will publicly release the code, data, and additional materials required to reproduce the experiments and results presented in this work.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12467–12479, 2025.

Yang Bai, Liudi Yang, George Eskandar, Fengyi Shen, Mohammad Altillawi, Ziyuan Liu, and Gitta Kutyniok. DRAW2ACT: Turning depth-encoded trajectories into robotic demonstration videos. *arXiv preprint arXiv:2512.14217*, 2025.

Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8709–8719, 2019.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision – ECCV 2020*, pp. 361–378. Springer, 2020.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts, 2025.

Matthew Chang, Aditya Prakash, and Saurabh Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, et al. Bridgev2w: Bridging video generation models to embodied world models via embodiment masks. *arXiv preprint arXiv:2602.03793*, 2026.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. *arXiv preprint arXiv:2512.08765*, 2025.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Kaelbling, et al. Video language planning. In *International Conference on Learning Representations*, volume 2024, pp. 31138–31155, 2024.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loïc Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. Dreamgen: Unlocking generalization in robot learning through video world models. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 5170–5194. PMLR, 2025. URL https://proceedings.mlr.press/v305/jang25a.html.

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2040–2049, 2025.

Tobias Jülg, Seongjin Bien, Simon Hilber, Yannik Blei, Pierre Krack, Maximilian Li, Sven Parusel, Rudolf Lioutikov, Florian Walter, and Wolfram Burgard. DuoBench: A reproducible benchmark for bimanual manipulation in simulation and the real world, 2026. URL https://arxiv.org/abs/2606.11901.

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*, 2024.

Mathis Koroglu, Hugo Caselles-Dupré, Guillaume Jeanneret, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6226–6236, 2025.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10138–10148, 2021.

Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6091–6099, 2026.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3943–3960. PMLR, 2025. URL https://proceedings.mlr.press/v270/liang25b.html.

Wenjun Lin, Jensen Zhang, Kaitong Cai, and Keze Wang. Storm: Search-guided generative world models for robotic manipulation. *arXiv preprint arXiv:2512.18477*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44776–44791, 2023.

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 (CVPR)*, pp. 21740–21751, 2024.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3d full-body poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4269–4279, June 2026.

Tao Pang, H. J. Terry Suh, Lujie Yang, and Russ Tedrake. Global planning for contact-rich manipulation via local smoothing of quasi-dynamic contact models. *IEEE Transactions on Robotics*, 39 (6):4691–4711, 2023.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9826–9836, 2024.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Karol Hausman, Brian Ichter, Sergey Levine, Karl Pertsch, Quan Vuong, Homer Walke, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*, 2022.

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.

Martin Sedlacek, Pavlo Yefanov, Georgy Ponimatkin, Jai Bardhan, Simon Pilc, Mederic Fourmy, Evangelos Kazakos, Cees G. M. Snoek, Josef Sivic, and Vladimir Petrik. REALM: A real-to-sim validated benchmark for generalization in robotic manipulation, 2025.

Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-graspnet: Efficient 6-DoF grasp generation in cluttered scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13438–13444, 2021.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Angtian Wang, Haibin Huang, Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: Any trajectory instruction for controllable video generation. *arXiv preprint*, arXiv:2505.22944, 2025b.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025c. URL https://openreview.net/forum?id=hpu6r8oLw9.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. In *Robotics: Science and Systems*, 2026. URL https://www.yixuanwang.me/interactive_world_sim/.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12713–12724, 2025d.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: Zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260, 2025.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. In *Robotics: Science and Systems (RSS)*, 2024.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NxoFmGgWC9.

Zhuoyuan Wu and Jun Gao. Oscar: Omni-embodiment action-conditioned world model for robotics, 2026. URL https://arxiv.org/abs/2606.04463.

Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. In *8th Annual Conference on Robot Learning (CoRL)*, 2024.

Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4d: Kinematic 4d world modeling for spatiotemporal embodied simulation. *arXiv preprint arXiv:2603.16669*, 2026.

Zhao-Heng Yin, Sherry Yang, and Pieter Abbeel. Object-centric 3d motion field for robot learning from human videos. *Advances in Neural Information Processing Systems*, 38:55923–55943, 2026.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27683–27693, 2025.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Hongyan Zhi, Peihao Chen, Siyuan Zhou, Yubo Dong, Quanxi Wu, Lei Han, and Mingkui Tan. 3DFlowAction: Learning cross-embodiment manipulation from 3d flow world model. *arXiv preprint arXiv:2506.06199*, 2025.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61885–61896. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhou24f.html.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844, October 2025.

A RESULTS PER EVALUATION SET

A.1 OUT-OF-DOMAIN HUMAN

Table 4: HO-Cap Wang et al. (2025c) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.453	0.463	0.194	0.001	22.63	0.865	0.103	0.053
	Contact pointmap video (ours)	0.485	0.502	0.191	0.005	21.69	0.843	0.118	0.056
	Contact latent transport (ours)	0.414	0.417	0.218	0.003	22.10	0.860	0.110	0.057
	Actor skeleton baseline	0.450	0.362	0.455	0.011	20.55	0.838	0.124	0.067
	Actor pointmap baseline	0.506	0.394	0.413	0.007	22.90	0.865	0.111	0.064
Sim	**Contact tokens (ours)**	0.382	0.351	0.292	0.014	21.30	0.844	0.128	0.073
	Contact pointmap video (ours)	0.405	0.331	0.348	0.026	20.78	0.822	0.143	0.079
	Contact latent transport (ours)	0.345	0.266	0.372	0.031	20.58	0.829	0.140	0.082
	Actor skeleton baseline	0.337	0.217	0.654	0.032	21.29	0.826	0.134	0.079
	Actor pointmap baseline	0.396	0.245	0.628	0.038	21.98	0.841	0.123	0.074
Human + Sim	**Contact tokens (ours)**	0.451	0.469	0.194	0.008	22.38	0.860	0.107	0.055
	Contact pointmap video (ours)	0.488	0.484	0.191	**0.000**	22.14	0.855	0.118	0.061
	Contact latent transport (ours)	0.413	0.406	0.237	0.013	22.13	0.858	0.110	0.058
	Actor skeleton baseline	0.445	0.318	0.518	0.017	22.77	0.857	0.102	0.054
	Actor pointmap baseline	0.527	0.399	0.455	0.022	22.78	0.858	0.103	0.057
Human + Sim + Robot	**Contact tokens (ours)**	0.445	0.461	**0.172**	**0.000**	22.61	0.862	0.104	0.054
	Contact pointmap video (ours)	0.486	0.502	0.174	0.003	22.18	0.853	0.110	0.053
	Contact latent transport (ours)	0.439	0.412	0.213	0.008	22.59	0.862	0.103	0.052
	Actor skeleton baseline	0.447	0.339	0.453	0.017	22.07	0.858	0.107	0.056
	Actor pointmap baseline	0.515	0.412	0.376	0.010	23.67	**0.866**	0.089	0.046
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.537	0.504	0.185	0.013	23.33	0.861	0.095	0.048
	+ actor pointmap	**0.572**	**0.537**	0.177	0.008	**23.68**	0.863	**0.086**	**0.044**
	w/o actor down-weighting	0.436	0.461	0.200	**0.000**	22.44	0.860	0.107	0.056
	w/o object up-weighting	0.426	0.428	0.237	0.008	22.35	0.861	0.106	0.053
	+ pre-contact loss	0.453	0.467	0.193	0.005	22.60	0.862	0.103	0.051
	+ depth head	0.432	0.415	0.243	0.006	22.05	0.840	0.122	0.067
	+ frame-0 object mask	0.469	0.474	0.194	**0.000**	22.63	0.863	0.103	0.052
	w/o frame-0 depth	0.462	0.473	0.186	0.003	22.59	0.862	0.103	0.052
	w/o slot state	0.438	0.448	0.220	0.008	22.56	0.862	0.105	0.054
External models	TesserAct Zhen et al. (2025)	0.324	0.272	0.577	**0.000**	20.22	0.831	0.160	0.094
	CTRL-World Guo et al. (2025)	0.295	0.245	0.765	0.059	18.38	0.710	0.407	0.237
	OSCAR Wu & Gao (2026)	0.317	0.236	0.635	0.001	21.87	0.834	0.170	0.096
	Kinema4D Xu et al. (2026)	0.306	0.203	0.766	0.035	22.63	0.860	0.123	0.080
	ATI Wang et al. (2025b)	0.434	0.398	0.347	0.045	14.58	0.579	0.452	0.208
	Wan-Move Chu et al. (2025)	0.433	0.384	0.274	0.009	14.90	0.575	0.400	0.159

Table 5: HOGraspNet results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.387	0.544	0.089	**0.000**	21.00	0.849	0.107	0.065
	Contact pointmap video (ours)	0.423	**0.618**	**0.080**	**0.000**	20.16	0.824	0.122	0.071
	Contact latent transport (ours)	0.398	0.568	0.091	**0.000**	21.20	0.854	0.104	0.067
	Actor skeleton baseline	0.426	0.570	0.093	0.014	18.54	0.775	0.149	0.091
	Actor pointmap baseline	0.419	0.557	0.117	0.002	21.67	0.854	0.098	0.066
Sim	**Contact tokens (ours)**	0.406	0.505	0.146	0.006	20.79	0.845	0.113	0.074
	Contact pointmap video (ours)	0.403	0.489	0.125	0.002	20.10	0.821	0.133	0.086
	Contact latent transport (ours)	0.396	0.472	0.191	0.028	20.44	0.837	0.122	0.078
	Actor skeleton baseline	0.413	0.473	0.170	0.001	19.55	0.802	0.143	0.093
	Actor pointmap baseline	0.413	0.441	0.181	0.005	20.43	0.835	0.115	0.086
Human + Sim	**Contact tokens (ours)**	0.411	0.557	0.093	**0.000**	21.15	0.849	0.106	0.066
	Contact pointmap video (ours)	0.454	0.559	0.094	**0.000**	21.31	0.851	0.106	0.065
	Contact latent transport (ours)	0.402	0.577	0.092	0.001	21.10	0.850	0.106	0.066
	Actor skeleton baseline	0.437	0.575	0.118	0.008	20.35	0.836	0.114	0.081
	Actor pointmap baseline	0.429	0.526	0.197	0.005	19.85	0.824	0.129	0.086
Human + Sim + Robot	**Contact tokens (ours)**	0.404	0.530	0.100	**0.000**	21.23	0.852	0.104	0.063
	Contact pointmap video (ours)	0.428	0.592	0.100	0.004	20.88	0.842	0.111	0.068
	Contact latent transport (ours)	0.410	0.569	0.083	0.013	21.40	0.855	0.101	**0.061**
	Actor skeleton baseline	0.420	0.570	0.091	0.003	20.01	0.836	0.121	0.080
	Actor pointmap baseline	0.443	0.548	0.108	0.017	**21.74**	0.850	**0.094**	0.064
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.450	0.594	0.096	0.002	20.77	0.835	0.113	0.074
	+ actor pointmap	**0.467**	0.574	0.116	0.004	21.15	0.838	0.104	0.072
	w/o actor down-weighting	0.396	0.544	0.106	0.002	21.16	0.851	0.106	0.067
	w/o object up-weighting	0.394	0.541	0.100	0.003	21.22	0.853	0.104	0.064
	+ pre-contact loss	0.405	0.550	0.090	0.004	21.26	0.855	0.103	0.065
	+ depth head	0.383	0.536	0.096	**0.000**	20.45	0.821	0.130	0.085
	+ frame-0 object mask	0.418	0.565	0.085	**0.000**	21.44	**0.856**	0.100	0.062
	w/o frame-0 depth	0.407	0.561	0.117	0.004	21.31	0.852	0.105	0.068
	w/o slot state	0.403	0.547	0.113	**0.000**	21.10	0.852	0.105	0.068
External models	TesserAct Zhen et al. (2025)	0.342	0.470	0.422	0.026	11.71	0.613	0.355	0.172
	CTRL-World Guo et al. (2025)	0.334	0.445	0.481	0.116	17.12	0.687	0.404	0.228
	OSCAR Wu & Gao (2026)	0.370	0.509	0.135	0.001	20.28	0.806	0.177	0.106
	Kinema4D Xu et al. (2026)	0.347	0.452	0.236	0.015	14.24	0.723	0.234	0.118
	ATI Wang et al. (2025b)	0.393	0.367	0.199	0.001	14.90	0.568	0.385	0.182
	Wan-Move Chu et al. (2025)	0.388	0.392	0.189	**0.000**	16.71	0.669	0.241	0.124

A.2 IN-DOMAIN ROBOT

Table 6: DROID Khazatsky et al. (2024) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.291	0.200	0.543	0.045	24.76	**0.887**	0.116	0.058
	Contact pointmap video (ours)	0.310	0.206	0.601	0.048	23.00	0.859	0.138	0.066
	Contact latent transport (ours)	0.299	0.178	0.663	0.083	24.43	0.881	0.118	0.063
	Actor skeleton baseline	0.233	0.110	1.051	0.083	19.69	0.768	0.188	0.098
	Actor pointmap baseline	0.274	0.133	0.906	0.085	23.38	0.865	0.128	0.072
Sim	**Contact tokens (ours)**	0.351	0.257	0.475	0.031	23.83	0.873	0.125	0.065
	Contact pointmap video (ours)	0.354	0.238	0.465	0.026	23.10	0.843	0.146	0.075
	Contact latent transport (ours)	0.316	0.216	0.539	0.041	23.62	0.867	0.127	0.065
	Actor skeleton baseline	0.356	0.231	0.661	0.056	22.78	0.836	0.133	0.074
	Actor pointmap baseline	0.384	0.253	0.633	0.057	23.99	0.871	0.109	0.062
Human + Sim	**Contact tokens (ours)**	0.350	0.259	0.431	0.020	24.08	0.877	0.118	0.063
	Contact pointmap video (ours)	0.363	0.245	0.441	0.025	23.83	0.872	0.125	0.065
	Contact latent transport (ours)	0.345	0.234	0.442	0.033	23.57	0.868	0.129	0.072
	Actor skeleton baseline	0.351	0.240	0.580	0.053	23.09	0.857	0.124	0.074
	Actor pointmap baseline	0.389	0.254	0.565	0.049	22.54	0.844	0.132	0.077
Human + Sim + Robot	**Contact tokens (ours)**	0.395	0.279	0.440	0.013	24.88	0.881	0.108	0.055
	Contact pointmap video (ours)	0.438	0.307	0.437	0.015	24.47	0.875	0.113	0.054
	Contact latent transport (ours)	0.396	0.273	0.421	**0.010**	24.95	0.882	0.108	0.053
	Actor skeleton baseline	0.417	0.311	0.452	0.031	23.55	0.869	0.108	0.061
	Actor pointmap baseline	0.440	0.309	0.446	0.024	24.18	0.875	0.102	0.057
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.451	0.325	0.396	0.023	24.21	0.867	0.103	0.056
	+ actor pointmap	**0.488**	**0.371**	**0.353**	0.017	24.52	0.877	**0.096**	0.053
	w/o actor down-weighting	0.406	0.292	0.429	0.014	24.92	0.883	0.106	**0.052**
	w/o object up-weighting	0.393	0.271	0.434	0.022	24.92	0.883	0.106	0.053
	+ pre-contact loss	0.392	0.270	0.438	0.016	24.84	0.881	0.108	0.054
	+ depth head	0.385	0.269	0.506	0.027	23.84	0.856	0.129	0.070
	+ frame-0 object mask	0.403	0.271	0.487	0.020	**24.97**	0.886	0.106	0.054
	w/o frame-0 depth	0.392	0.263	0.457	0.013	24.93	0.884	0.107	0.053
	w/o slot state	0.395	0.289	0.474	0.024	24.69	0.880	0.109	0.056
External models	TesserAct Zhen et al. (2025)	0.215	0.113	1.029	0.033	13.42	0.644	0.343	0.153
	CTRL-World Guo et al. (2025)	0.183	0.076	1.346	0.192	21.48	0.758	0.308	0.157
	OSCAR Wu & Gao (2026)	0.346	0.220	0.593	0.014	24.41	0.862	0.138	0.074
	Kinema4D Xu et al. (2026)	0.282	0.160	0.847	0.041	16.15	0.753	0.211	0.095
	ATI Wang et al. (2025b)	0.265	0.157	0.712	0.080	13.11	0.492	0.544	0.298
	Wan-Move Chu et al. (2025)	0.259	0.146	0.776	0.029	16.02	0.600	0.361	0.141

Table 7: RH20T val Fang et al. (2024) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.279	0.209	0.601	0.090	23.93	0.862	0.129	0.066
	Contact pointmap video (ours)	0.314	0.222	0.614	0.063	22.84	0.836	0.140	0.073
	Contact latent transport (ours)	0.272	0.190	0.727	0.145	23.54	0.855	0.130	0.071
	Actor skeleton baseline	0.235	0.153	1.075	0.282	13.34	0.507	0.374	0.234
	Actor pointmap baseline	0.219	0.089	1.094	0.055	23.87	0.861	0.136	0.074
Sim	**Contact tokens (ours)**	0.349	0.263	0.516	0.032	23.65	0.859	0.129	0.072
	Contact pointmap video (ours)	0.328	0.202	0.718	0.098	21.61	0.753	0.224	0.123
	Contact latent transport (ours)	0.306	0.202	0.645	0.054	23.44	0.849	0.136	0.072
	Actor skeleton baseline	0.308	0.200	0.919	0.180	18.43	0.723	0.283	0.172
	Actor pointmap baseline	0.196	0.089	1.092	0.112	22.27	0.826	0.163	0.089
Human + Sim	**Contact tokens (ours)**	0.349	0.253	0.500	0.033	23.49	0.848	0.134	0.073
	Contact pointmap video (ours)	0.362	0.260	0.495	0.024	23.40	0.847	0.143	0.079
	Contact latent transport (ours)	0.342	0.236	0.511	0.059	23.49	0.850	0.134	0.073
	Actor skeleton baseline	0.305	0.172	0.946	0.154	19.85	0.789	0.206	0.115
	Actor pointmap baseline	0.126	0.050	1.397	0.318	18.16	0.743	0.244	0.154
Human + Sim + Robot	**Contact tokens (ours)**	0.387	0.285	**0.397**	0.017	23.89	0.859	0.120	0.065
	Contact pointmap video (ours)	0.403	0.284	0.523	0.044	23.62	0.846	0.123	0.066
	Contact latent transport (ours)	0.384	0.252	0.490	0.017	**24.45**	**0.866**	**0.113**	**0.058**
	Actor skeleton baseline	0.393	0.303	0.658	0.103	19.65	0.790	0.203	0.118
	Actor pointmap baseline	0.167	0.086	1.090	0.031	22.81	0.849	0.142	0.078
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	**0.442**	**0.342**	0.445	0.072	21.74	0.815	0.165	0.092
	+ actor pointmap	0.375	0.250	0.500	0.006	23.64	0.852	0.127	0.066
	w/o actor down-weighting	0.388	0.278	0.456	0.005	23.79	0.860	0.123	0.066
	w/o object up-weighting	0.382	0.271	0.505	0.052	24.11	0.861	0.116	0.063
	+ pre-contact loss	0.385	0.268	0.455	0.016	24.16	0.864	0.116	0.062
	+ depth head	0.374	0.266	0.490	0.034	22.51	0.821	0.151	0.081
	+ frame-0 object mask	0.392	0.286	0.512	0.041	24.13	0.861	0.120	0.065
	w/o frame-0 depth	0.396	0.286	0.467	0.031	24.34	0.865	0.116	0.061
	w/o slot state	0.386	0.273	0.476	0.019	23.90	0.862	0.119	0.065
External models	TesserAct Zhen et al. (2025)	0.206	0.081	1.107	0.087	11.55	0.594	0.394	0.209
	CTRL-World Guo et al. (2025)	0.213	0.090	1.197	0.142	20.37	0.674	0.334	0.178
	OSCAR Wu & Gao (2026)	0.277	0.164	0.628	**0.000**	24.43	0.853	0.147	0.080
	Kinema4D Xu et al. (2026)	0.317	0.233	0.821	0.087	14.50	0.718	0.249	0.115
	ATI Wang et al. (2025b)	0.296	0.177	0.675	0.070	14.21	0.456	0.517	0.300
	Wan-Move Chu et al. (2025)	0.275	0.151	0.634	0.041	17.26	0.575	0.323	0.145

Table 8: RoboMIND val results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.224	0.068	1.126	0.032	24.94	0.882	0.184	0.081
	Contact pointmap video (ours)	0.246	0.089	0.982	0.023	22.90	0.857	0.212	0.096
	Contact latent transport (ours)	0.219	0.059	1.110	0.063	24.58	0.877	0.187	0.085
	Actor skeleton baseline	0.170	0.038	1.473	0.278	16.25	0.684	0.338	0.200
	Actor pointmap baseline	0.192	0.053	1.461	0.042	24.11	0.866	0.225	0.107
Sim	**Contact tokens (ours)**	0.216	0.070	0.979	0.062	24.55	0.874	0.193	0.091
	Contact pointmap video (ours)	0.210	0.072	1.007	0.105	23.77	0.847	0.229	0.121
	Contact latent transport (ours)	0.222	0.079	1.026	0.058	24.29	0.865	0.196	0.090
	Actor skeleton baseline	0.233	0.109	1.106	0.137	21.73	0.818	0.249	0.153
	Actor pointmap baseline	0.180	0.065	1.353	0.042	23.85	0.853	0.223	0.091
Human + Sim	**Contact tokens (ours)**	0.215	0.062	1.110	0.033	24.59	0.877	0.195	0.089
	Contact pointmap video (ours)	0.261	0.128	0.807	0.066	24.38	0.872	0.184	0.093
	Contact latent transport (ours)	0.221	0.072	0.982	0.033	24.55	0.874	0.190	0.090
	Actor skeleton baseline	0.253	0.100	1.175	0.098	21.89	0.838	0.213	0.106
	Actor pointmap baseline	0.130	0.018	1.623	0.090	22.43	0.836	0.282	0.134
Human + Sim + Robot	**Contact tokens (ours)**	0.302	0.168	0.691	0.030	25.07	0.882	0.165	0.072
	Contact pointmap video (ours)	0.342	0.192	0.663	**0.007**	23.41	0.865	0.168	0.078
	Contact latent transport (ours)	0.302	0.143	0.817	0.022	25.17	0.885	0.163	0.071
	Actor skeleton baseline	0.323	0.234	0.713	0.116	22.41	0.866	0.163	0.097
	Actor pointmap baseline	0.171	0.060	1.250	0.056	23.69	0.869	0.209	0.091
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	**0.393**	**0.278**	**0.614**	0.056	23.96	0.871	**0.145**	0.080
	+ actor pointmap	0.295	0.155	0.727	0.021	24.75	0.879	0.169	0.071
	w/o actor down-weighting	0.310	0.141	0.797	0.010	24.97	0.882	0.163	0.073
	w/o object up-weighting	0.299	0.133	0.827	0.028	**25.31**	**0.886**	0.162	**0.069**
	+ pre-contact loss	0.312	0.158	0.707	0.028	24.92	0.882	0.165	0.073
	+ depth head	0.298	0.167	0.729	0.024	24.05	0.866	0.177	0.086
	+ frame-0 object mask	0.325	0.173	0.713	0.025	25.05	0.884	0.163	0.071
	w/o frame-0 depth	0.328	0.170	0.717	0.013	25.15	0.883	0.167	0.074
	w/o slot state	0.308	0.167	0.720	0.023	24.90	0.883	0.164	0.072
External models	TesserAct Zhen et al. (2025)	0.176	0.039	1.445	0.063	13.94	0.689	0.380	0.195
	CTRL-World Guo et al. (2025)	0.176	0.024	1.558	0.131	21.67	0.792	0.350	0.185
	OSCAR Wu & Gao (2026)	0.354	0.246	0.638	**0.007**	24.49	0.868	0.192	0.110
	Kinema4D Xu et al. (2026)	0.293	0.164	0.749	0.069	16.87	0.794	0.214	0.122
	ATI Wang et al. (2025b)	0.183	0.024	1.277	0.019	17.79	0.699	0.469	0.245
	Wan-Move Chu et al. (2025)	0.182	0.033	1.332	0.046	19.00	0.735	0.335	0.133

A.3 Out-of-domain real robot

Table 9: DuoBench real Jülg et al. (2026) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.238	0.173	0.798	0.097	25.63	0.906	0.112	0.027
	Contact pointmap video (ours)	0.271	0.164	0.858	0.109	23.98	0.879	0.128	0.036
	Contact latent transport (ours)	0.232	0.122	0.957	0.135	26.22	0.907	0.105	0.024
	Actor skeleton baseline	0.135	0.029	1.635	0.245	19.15	0.800	0.174	0.081
	Actor pointmap baseline	0.158	0.047	1.476	0.268	23.79	0.880	0.121	0.039
Sim	**Contact tokens (ours)**	**0.372**	0.240	0.532	0.008	25.66	0.902	0.103	0.032
	Contact pointmap video (ours)	0.340	0.231	0.613	0.059	25.34	0.889	0.114	0.037
	Contact latent transport (ours)	0.327	0.205	0.652	0.045	25.71	0.898	0.103	0.029
	Actor skeleton baseline	0.149	0.052	1.410	0.352	23.85	0.865	0.117	0.043
	Actor pointmap baseline	0.186	0.093	1.521	0.491	24.29	0.880	0.103	0.035
Human + Sim	**Contact tokens (ours)**	0.359	0.247	0.484	0.018	26.33	0.903	0.097	0.027
	Contact pointmap video (ours)	0.366	0.252	0.497	0.014	26.17	0.902	0.099	0.025
	Contact latent transport (ours)	0.331	0.192	0.598	0.023	26.05	0.900	0.100	0.028
	Actor skeleton baseline	0.176	0.074	1.258	0.247	24.21	0.880	0.107	0.037
	Actor pointmap baseline	0.170	0.077	1.524	0.431	23.08	0.864	0.125	0.054
Human + Sim + Robot	**Contact tokens (ours)**	0.357	0.252	0.476	**0.000**	26.87	0.906	0.094	**0.019**
	Contact pointmap video (ours)	0.368	0.255	0.533	0.008	26.50	0.902	0.098	0.022
	Contact latent transport (ours)	0.351	0.249	0.595	0.015	26.72	0.905	0.095	0.020
	Actor skeleton baseline	0.220	0.091	1.119	0.252	20.20	0.861	0.152	0.083
	Actor pointmap baseline	0.231	0.128	1.213	0.286	24.87	0.885	0.095	0.030
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.315	0.126	0.833	0.112	23.87	0.875	0.105	0.038
	+ actor pointmap	0.296	0.140	1.158	0.244	25.48	0.889	**0.087**	0.026
	w/o actor down-weighting	0.353	0.256	0.521	0.010	26.96	0.908	0.093	**0.019**
	w/o object up-weighting	0.356	**0.259**	**0.473**	0.004	27.02	0.907	0.092	**0.019**
	+ pre-contact loss	0.339	0.222	0.515	0.004	26.87	0.907	0.093	**0.019**
	+ depth head	0.318	0.205	0.757	0.067	24.79	0.873	0.117	0.040
	+ frame-0 object mask	0.345	0.228	0.578	0.006	26.95	0.907	0.093	**0.019**
	w/o frame-0 depth	0.346	0.246	0.509	0.015	26.73	0.905	0.095	0.020
	w/o slot state	0.355	0.255	0.514	0.003	**27.06**	**0.909**	0.092	**0.019**
External models	TesserAct Zhen et al. (2025)	0.200	0.047	1.804	0.010	13.07	0.671	0.340	0.190
	CTRL-World Guo et al. (2025)	0.135	0.040	1.824	0.275	21.65	0.760	0.308	0.141
	OSCAR Wu & Gao (2026)	0.320	0.191	0.809	0.060	24.88	0.853	0.134	0.063
	Kinema4D Xu et al. (2026)	0.263	0.142	1.010	0.154	15.94	0.739	0.223	0.111
	ATI Wang et al. (2025b)	0.325	0.152	0.632	0.007	15.82	0.615	0.459	0.208
	Wan-Move Chu et al. (2025)	0.288	0.153	0.916	0.045	21.42	0.756	0.217	0.065

Table 10: AgiBot Bu et al. (2025) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.245	0.152	0.666	0.015	24.97	0.890	0.137	0.082
	Contact pointmap video (ours)	0.250	0.159	0.641	0.025	22.76	0.861	0.169	0.081
	Contact latent transport (ours)	0.235	0.140	0.659	0.036	24.91	0.889	0.137	0.079
	Actor skeleton baseline	0.144	0.060	1.359	0.403	13.54	0.584	0.391	0.230
	Actor pointmap baseline	0.115	0.019	1.378	0.095	23.70	0.876	0.177	0.094
Sim	**Contact tokens (ours)**	0.252	0.188	0.501	0.032	25.30	0.889	0.139	0.065
	Contact pointmap video (ours)	0.262	0.190	0.532	0.053	24.08	0.856	0.179	0.089
	Contact latent transport (ours)	0.273	0.186	0.532	0.030	25.02	0.884	0.145	**0.061**
	Actor skeleton baseline	0.217	0.160	1.055	0.249	21.64	0.828	0.214	0.099
	Actor pointmap baseline	0.084	0.012	1.444	0.195	23.39	0.857	0.196	0.087
Human + Sim	**Contact tokens (ours)**	0.269	0.213	0.472	0.014	**25.90**	**0.895**	0.132	0.062
	Contact pointmap video (ours)	0.308	0.249	0.393	**0.005**	25.06	0.886	0.141	0.067
	Contact latent transport (ours)	0.271	0.209	0.535	0.046	25.33	0.890	0.132	0.071
	Actor skeleton baseline	0.222	0.166	0.864	0.181	22.71	0.863	0.161	0.083
	Actor pointmap baseline	0.079	0.010	1.538	0.354	20.09	0.832	0.248	0.148
Human + Sim + Robot	**Contact tokens (ours)**	0.294	0.234	0.399	0.011	24.62	0.882	0.144	0.074
	Contact pointmap video (ours)	0.297	0.238	0.447	0.006	24.76	0.879	0.143	0.067
	Contact latent transport (ours)	0.300	0.255	0.468	0.022	25.42	0.890	**0.130**	0.067
	Actor skeleton baseline	0.280	0.225	0.620	0.091	21.98	0.865	0.168	0.098
	Actor pointmap baseline	0.108	0.017	1.417	0.069	22.99	0.865	0.196	0.089
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	**0.323**	**0.300**	0.414	0.062	23.32	0.863	0.152	0.080
	+ actor pointmap	0.273	0.195	0.480	0.017	24.91	0.886	0.142	0.066
	w/o actor down-weighting	0.276	0.220	0.443	0.020	24.53	0.883	0.147	0.074
	w/o object up-weighting	0.286	0.242	0.395	0.018	24.49	0.879	0.157	0.081
	+ pre-contact loss	0.285	0.210	0.468	0.042	24.47	0.880	0.155	0.083
	+ depth head	0.284	0.237	0.453	0.048	24.02	0.865	0.156	0.084
	+ frame-0 object mask	0.310	0.258	**0.387**	0.034	25.18	0.892	0.131	0.071
	w/o frame-0 depth	0.286	0.225	0.486	0.020	24.89	0.888	0.141	0.076
	w/o slot state	0.278	0.230	0.479	0.033	24.51	0.884	0.147	0.080
External models	TesserAct Zhen et al. (2025)	0.127	0.011	1.331	0.030	12.70	0.686	0.341	0.151
	CTRL-World Guo et al. (2025)	0.126	0.014	1.525	0.210	21.81	0.805	0.329	0.167
	OSCAR Wu & Gao (2026)	0.145	0.048	1.112	0.111	20.31	0.747	0.303	0.130
	Kinema4D Xu et al. (2026)	0.217	0.116	0.810	0.044	16.91	0.815	0.204	0.083
	ATI Wang et al. (2025b)	0.202	0.112	0.737	0.055	16.21	0.618	0.512	0.281
	Wan-Move Chu et al. (2025)	0.220	0.130	0.698	0.014	18.84	0.717	0.330	0.125

A.4 Out-of-domain simulator

Table 11: REALM Sedlacek et al. (2025) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.337	0.157	0.543	0.024	29.88	0.941	0.065	0.025
	Contact pointmap video (ours)	0.358	0.139	0.630	0.032	26.73	0.915	0.091	0.036
	Contact latent transport (ours)	0.346	0.174	0.535	0.018	30.46	0.942	0.065	0.022
	Actor skeleton baseline	0.298	0.092	0.764	**0.000**	23.81	0.888	0.133	0.069
	Actor pointmap baseline	0.357	0.139	0.585	0.001	28.52	0.932	0.074	0.027
Sim	**Contact tokens (ours)**	0.364	0.176	0.503	0.003	30.21	0.940	0.063	**0.018**
	Contact pointmap video (ours)	0.400	0.185	0.598	0.045	27.22	0.885	0.110	0.039
	Contact latent transport (ours)	0.295	0.095	0.714	**0.000**	29.87	0.939	0.064	0.021
	Actor skeleton baseline	0.403	0.182	0.561	0.005	28.73	0.929	0.056	0.023
	Actor pointmap baseline	0.401	0.152	0.601	0.018	29.82	0.942	**0.048**	0.021
Human + Sim	**Contact tokens (ours)**	0.393	0.201	0.472	0.006	30.34	0.941	0.063	0.019
	Contact pointmap video (ours)	0.416	0.225	0.433	0.003	29.40	0.935	0.065	0.021
	Contact latent transport (ours)	0.364	0.155	0.510	**0.000**	29.37	0.925	0.076	0.025
	Actor skeleton baseline	0.389	0.182	0.571	0.002	29.19	0.933	0.054	0.023
	Actor pointmap baseline	0.448	0.197	0.449	0.003	28.79	0.929	0.054	0.023
Human + Sim + Robot	**Contact tokens (ours)**	0.401	0.206	0.495	0.002	30.71	0.945	0.056	**0.018**
	Contact pointmap video (ours)	0.417	0.210	0.551	0.017	30.27	0.939	0.058	0.021
	Contact latent transport (ours)	0.384	0.184	0.548	0.003	30.33	0.942	0.056	0.019
	Actor skeleton baseline	0.447	0.206	0.565	0.032	28.86	0.936	0.051	0.021
	Actor pointmap baseline	0.393	0.149	0.671	0.011	29.78	0.943	**0.048**	0.020
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.454	**0.237**	0.488	0.013	28.79	0.928	0.054	0.022
	+ actor pointmap	**0.466**	0.207	0.471	0.003	30.19	0.943	**0.048**	0.021
	w/o actor down-weighting	0.405	0.203	0.441	0.003	**30.92**	0.945	0.054	**0.018**
	w/o object up-weighting	0.364	0.170	0.481	**0.000**	30.40	0.943	0.057	0.019
	+ pre-contact loss	0.387	0.205	0.505	0.005	30.81	0.945	0.055	**0.018**
	+ depth head	0.376	0.206	0.523	0.002	28.97	0.925	0.070	0.033
	+ frame-0 object mask	0.396	0.201	0.531	0.007	30.84	**0.946**	0.054	**0.018**
	w/o frame-0 depth	0.374	0.181	0.546	0.008	30.62	0.945	0.056	0.019
	w/o slot state	0.413	0.203	**0.424**	0.001	30.77	0.945	0.056	**0.018**
External models	TesserAct Zhen et al. (2025)	0.238	0.061	1.063	0.050	28.58	0.921	0.115	0.047
	CTRL-World Guo et al. (2025)	0.159	0.011	1.675	0.460	17.31	0.690	0.446	0.317
	OSCAR Wu & Gao (2026)	0.268	0.082	0.778	0.001	22.71	0.832	0.171	0.082
	Kinema4D Xu et al. (2026)	0.254	0.087	0.830	0.026	29.25	0.937	0.055	0.038
	ATI Wang et al. (2025b)	0.253	0.106	1.100	0.227	9.13	0.504	0.738	0.567
	Wan-Move Chu et al. (2025)	0.308	0.160	0.589	0.056	17.58	0.691	0.274	0.082

Table 12: DuoBench sim Jülg et al. (2026) results.

Training data	Model	IoU↑	End IoU↑	Final-pos. err.↓	Missing↓	PSNR↑	SSIM↑	LPIPS↓	DreamSim↓
Human	**Contact tokens (ours)**	0.198	0.145	0.720	0.141	16.20	0.795	0.270	0.119
	Contact pointmap video (ours)	0.225	0.133	0.774	0.206	16.23	0.784	0.273	0.126
	Contact latent transport (ours)	0.203	0.123	0.666	0.183	16.12	0.784	0.283	0.129
	Actor skeleton baseline	0.142	0.059	1.301	0.393	15.82	0.762	0.282	0.150
	Actor pointmap baseline	0.153	0.046	1.444	0.467	15.81	0.772	0.283	0.157
Sim	**Contact tokens (ours)**	0.241	0.200	0.571	0.143	16.24	0.780	0.277	0.136
	Contact pointmap video (ours)	0.264	0.171	0.688	0.148	16.15	0.773	0.283	0.137
	Contact latent transport (ours)	0.215	0.160	0.652	0.158	15.93	0.770	0.292	0.146
	Actor skeleton baseline	0.210	0.121	0.916	0.206	16.92	0.793	0.225	0.115
	Actor pointmap baseline	0.218	0.096	1.072	0.242	**17.18**	0.802	0.206	0.105
Human + Sim	**Contact tokens (ours)**	**0.278**	0.192	0.672	0.145	16.34	0.787	0.272	0.130
	Contact pointmap video (ours)	0.243	0.172	0.653	0.154	16.29	0.788	0.273	0.133
	Contact latent transport (ours)	0.209	0.136	0.795	0.207	15.55	0.762	0.301	0.159
	Actor skeleton baseline	0.205	0.087	0.993	0.218	16.40	0.784	0.240	0.126
	Actor pointmap baseline	0.211	0.097	1.143	0.233	16.75	0.789	0.222	0.124
Human + Sim + Robot	**Contact tokens (ours)**	0.248	0.210	0.529	0.083	16.12	0.781	0.279	0.134
	Contact pointmap video (ours)	0.262	0.229	0.460	0.072	16.15	0.779	0.275	0.137
	Contact latent transport (ours)	0.248	**0.244**	0.442	0.085	16.17	0.786	0.274	0.133
	Actor skeleton baseline	0.196	0.064	1.139	0.245	16.92	0.796	0.226	0.116
	Actor pointmap baseline	0.197	0.096	1.107	0.249	16.88	0.800	0.210	0.110
Ablations of contact tokens (Human + Sim + Robot)	+ actor skeleton	0.222	0.122	0.991	0.235	16.92	0.801	0.221	0.107
	+ actor pointmap	0.233	0.149	0.930	0.225	**17.18**	**0.805**	**0.203**	**0.103**
	w/o actor down-weighting	0.236	**0.244**	0.518	0.121	16.03	0.778	0.280	0.131
	w/o object up-weighting	0.242	0.186	0.561	0.080	16.06	0.778	0.285	0.139
	+ pre-contact loss	0.242	0.211	0.547	0.111	16.07	0.782	0.278	0.135
	+ depth head	0.226	0.176	0.824	0.195	16.13	0.753	0.274	0.155
	+ frame-0 object mask	0.247	0.211	0.511	0.090	16.32	0.787	0.272	0.130
	w/o frame-0 depth	0.271	0.225	**0.415**	**0.043**	16.22	0.786	0.273	0.132
	w/o slot state	0.232	0.208	0.568	0.096	16.15	0.787	0.276	0.133
External models	TesserAct Zhen et al. (2025)	0.142	0.109	0.821	0.050	16.15	0.795	0.285	0.139
	CTRL-World Guo et al. (2025)	0.090	0.016	1.660	0.486	13.57	0.475	0.561	0.351
	OSCAR Wu & Gao (2026)	0.151	0.104	0.983	0.140	14.77	0.632	0.312	0.183
	Kinema4D Xu et al. (2026)	0.138	0.101	1.065	0.137	16.16	0.795	0.245	0.125
	ATI Wang et al. (2025b)	0.182	0.166	0.628	0.126	8.56	0.383	0.727	0.575
	Wan-Move Chu et al. (2025)	0.154	0.118	0.712	0.102	13.70	0.719	0.351	0.166

B Dataset-Specific Preprocessing

B.1 Source Datasets

ContactFlow-$212K$ combines human, real-robot, and simulated robot interactions. For human interactions, we use TACO Liu et al. (2024) and OakInk Zhan et al. (2024), which provide multiview hand-object interactions with annotated 3D geometry, fixed-view egocentric videos from Taste-ROB Zhao et al. (2025), and smaller datasets with ground-truth contact annotations including HOI4D and ARCTIC. For real-robot interactions, we use DROID Khazatsky et al. (2024), RH20T Fang et al. (2024), RoboMind, and AgiBot Bu et al. (2025). We further include simulated interactions from LIBERO Liu et al. (2023), RoboTwin, RLBench James et al. (2020), Colosseum, DexMimicGen, robomimic, RMBench, RoboCasa, and GR-1 Tabletop Tasks.

B.2 Hand and Robot Geometry Estimation

Human hand estimation. We initialize the MANO Romero et al. (2017) hand model from the pose, shape, and translation predicted by HaMeR Pavlakos et al. (2024). Since independent predictions can vary across frames, we fix the hand shape to the median estimate over each sequence and jointly optimize pose and translation with Adam. The optimization fits the projected hand joints to the observed 2D keypoints and constrains the projected mesh with the hand mask. We regularize pose and translation over time to suppress frame-wise estimation noise. We keep scene depth separate from this optimization because monocular depth errors could otherwise distort the hand reconstruction. We first stabilize the scene depth using static background regions and then use the fixed MANO geometry as a metric reference to recover the scene scale. Finally, we preserve the hand dimensions and refine only its translation along the camera depth axis to align the hand surface with the reconstructed scene. The resulting geometry provides a metrically aligned hand surface for Contact Flow estimation.

Robot gripper pose estimation. Extracting accurate Contact Flow requires precise alignment between the robot geometry and the reconstructed scene. We segment the robot with SAM3 Carion et al. (2025) and reconstruct its articulated geometry from the recorded joint states using forward kinematics. We initialize the camera calibration from dataset annotations when available and otherwise evaluate multiple candidate transformations. We then render the robot geometry and compare its depth and silhouette with the observations. If the initial calibration is insufficient, we refine the camera-to-robot transformation while keeping the recorded joint states fixed. We minimize geometric depth and silhouette errors with a robust Levenberg-Marquardt solver and evaluate multiple initial transformations to reduce sensitivity to local minima. We retain the refined calibration only when the observed robot geometry provides sufficient support. The resulting alignment provides the metric gripper poses used to extract Contact Flow.

Dataset	Overall samples	Used for training
Human		
TASTE-Rob	35,870	35,611
ACE	20,506	17,247
TACO	13,654	13,284
OakInk2	12,033	8,853
ARCTIC	8,130	7,437
OakInk	3,024	3,015
Simulation		
RoboCasa	85,729	46,154
MimicGen	58,937	41,267
DexMimicGen	21,602	16,518
RLBench	18,026	10,766
ABCSim	9,786	8,742
LIBERO	6,360	6,109
ManiSkill	5,688	4,516
RoboTwin	4,073	2,274
RMBench	2,111	1,725
robomimic	1,963	1,151
Total	307,492	224,669

Table 13: Training data. A sample is a 49-frame clip with its contact flow. Samples are kept when they have at least one contact point, the object is visible in every frame and moves at least 5 px, and the camera is still, a few held-out tasks and embodiments are removed.