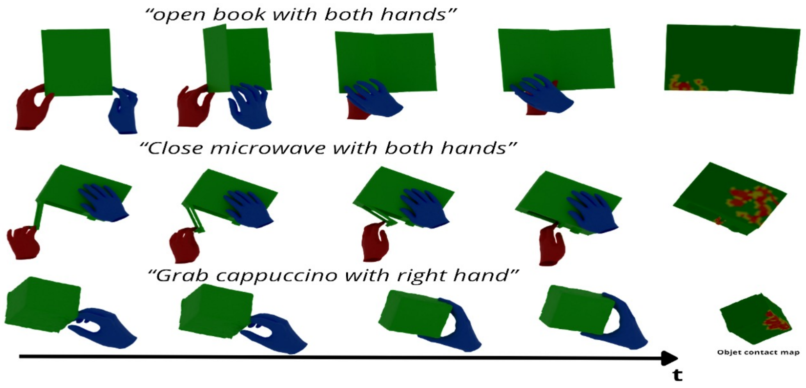
GCD: Geometry-Constrained Contact-Aware Diffusion for Text-Driven 3D Hand-Object Motion Synthesis



Josué kossi Adossehoun and Hazem Wannous

IMT Nord Europe, CRIStAL UMR 9189

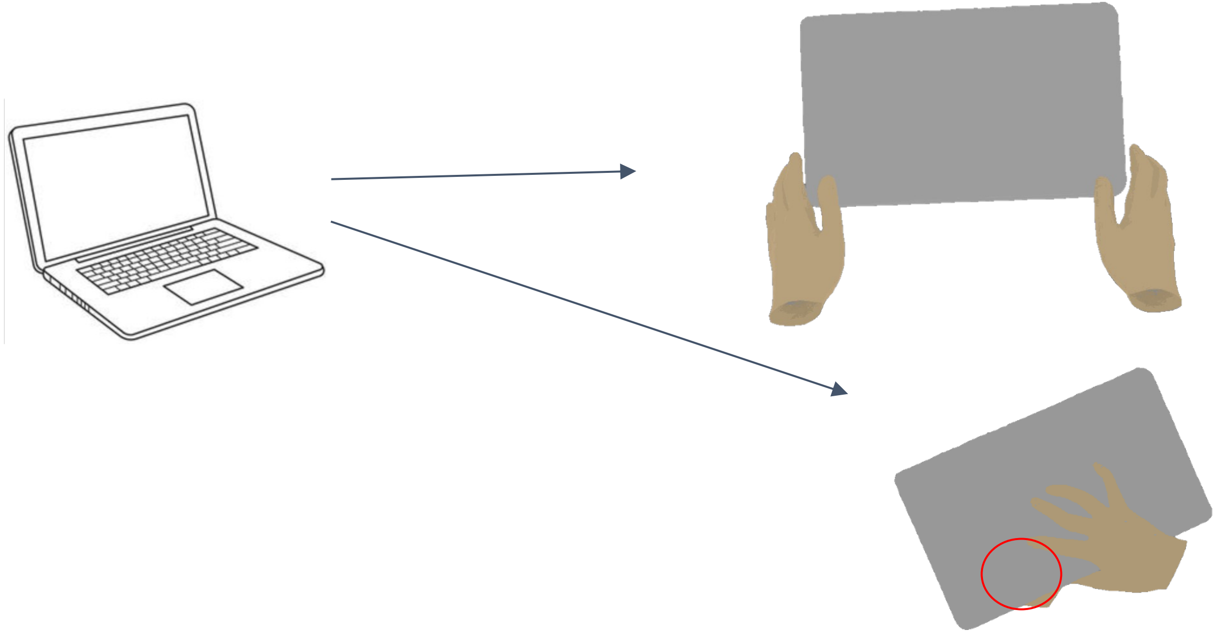
Lille, France

Overview

Generating 3D hand-object interaction motion from text and object geometry

Moravec's paradox — Easy to compute, hard to manipulate

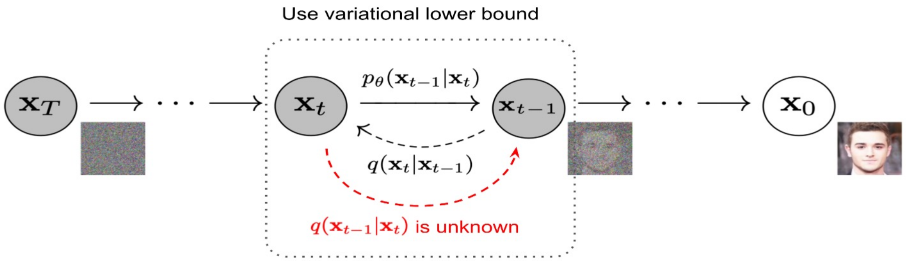
Open Laptop
with both hands +

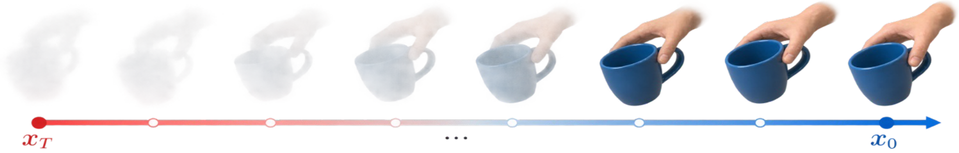


- Semantic alignment
- Physical plausibility

Motivation

Diffusion : a strong generative prior



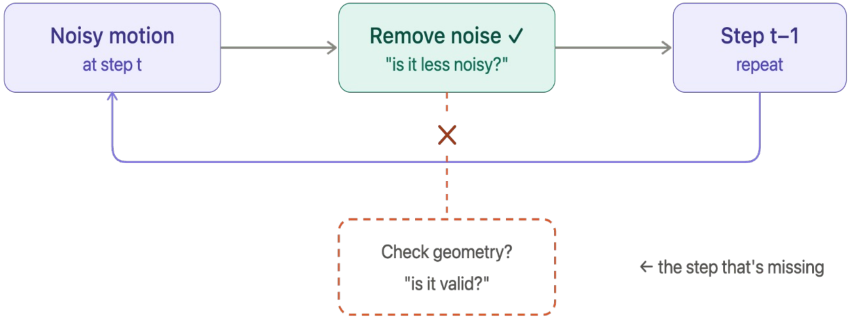


For **motion generation**, diffusion iteratively denoises an entire motion sequence while respecting temporal structure and conditioning cues.

Problem : learns the data distribution — not the laws of physics

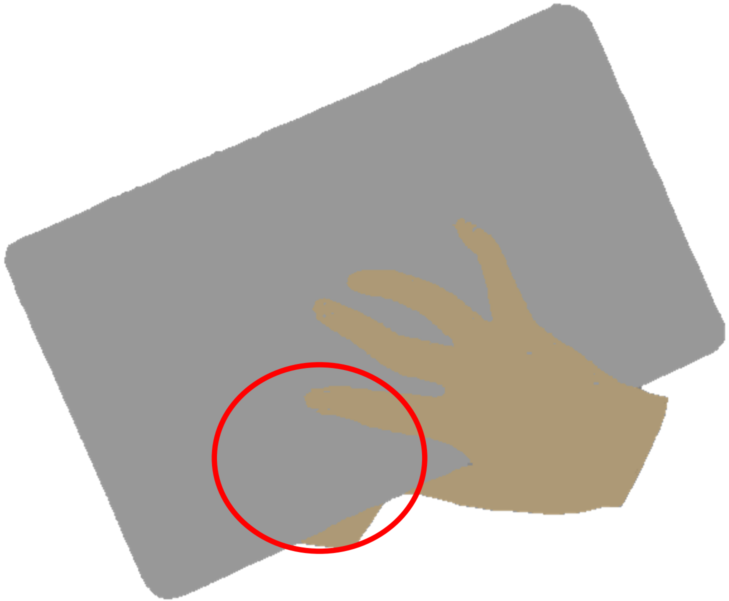
trajectory drifts,hand penetration, unstable grasps, or physically inconsistent interactions.

corrected only afterwards via post processing methods



Motivation

Key observation :

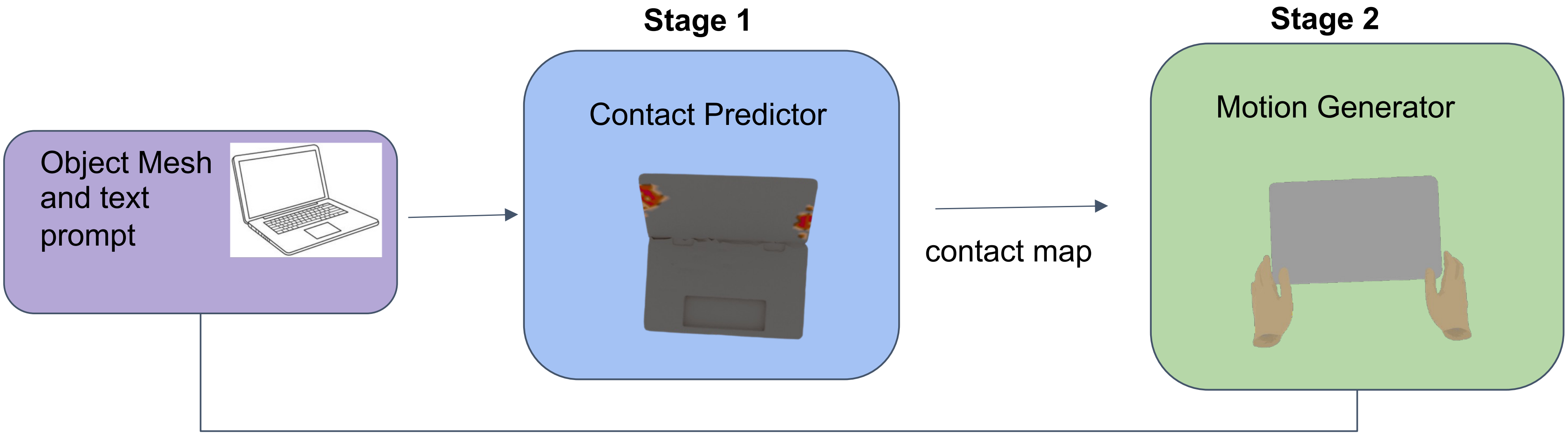


Hands and objects play asymmetric roles
- The object follows text intent
- The hand adapts to geometry

Method Overview

Design asymmetric dual-stream architecture.

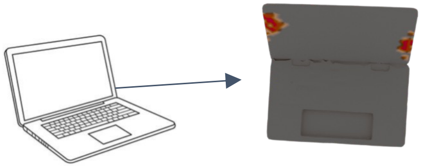
GCD first predicts where the interaction should occur, then generates how the hand and object move while respecting the underlying geometry.

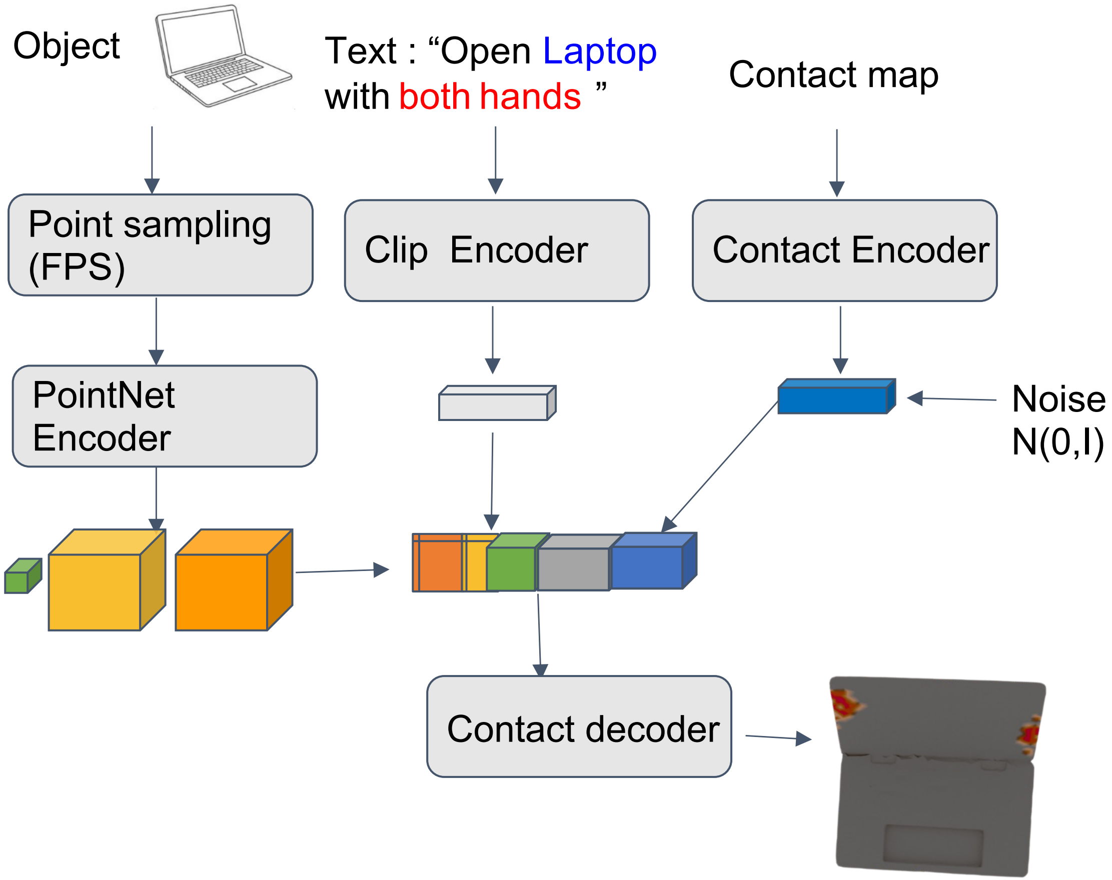


Stage 1 : contact predictor

Following text2hoi CVAE

Open Laptop
with both hands
+



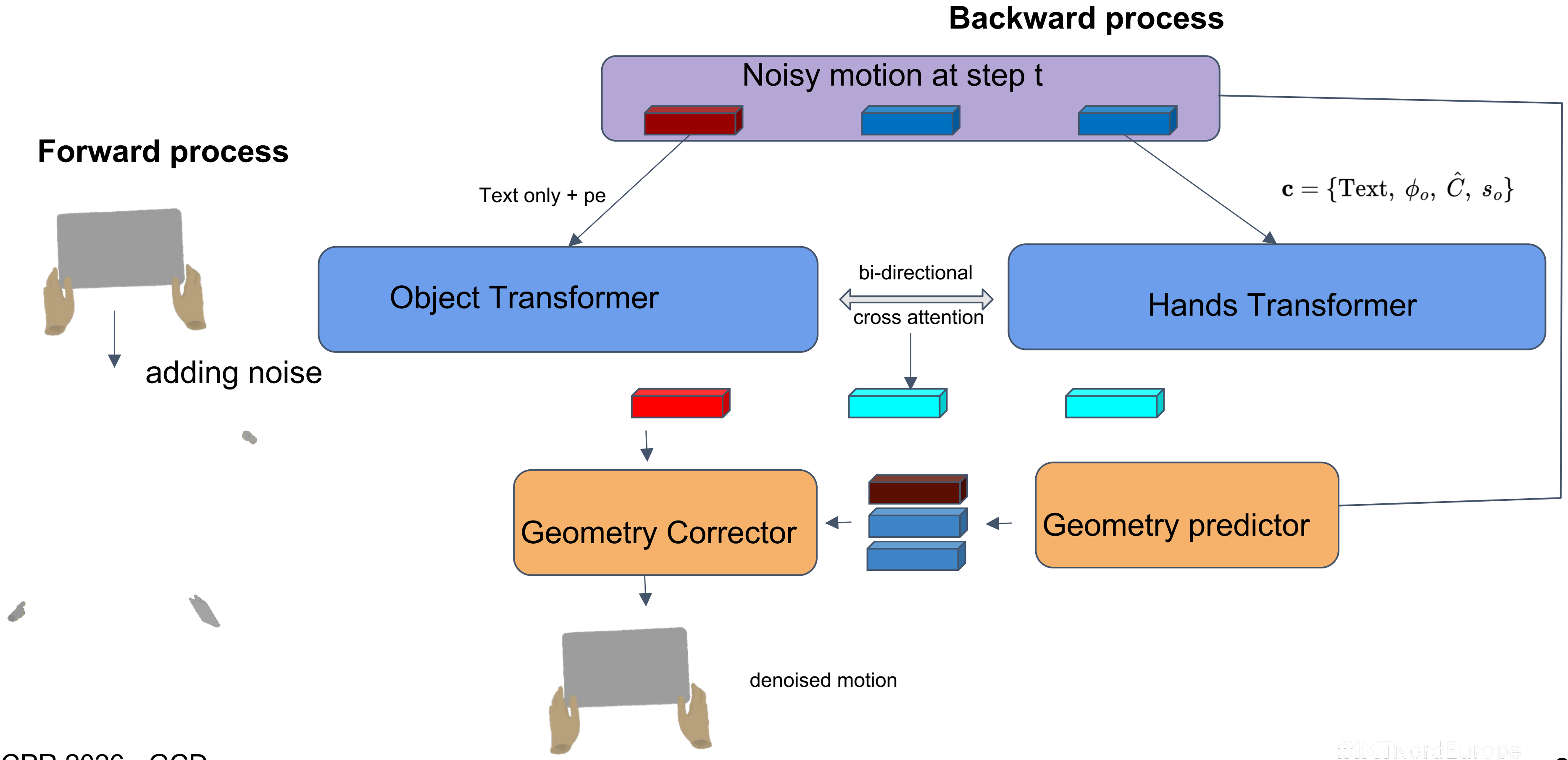


Supervision :

$$\mathcal{L}_{\text{contact}} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}.$$

- BCE: point-wise contact classification.
- Dice: alleviates class imbalance and improves mask quality.
- KL: regularizes the latent space for diverse contact generation.

Stage 2 : Motion generation



Training objectives

$$\mathcal{L}_{\mathrm{GCD}}(D_\theta) = \mathcal{L}_{\mathrm{diff}}(D_\theta) + \mathcal{L}_{\mathrm{dm}}(D_\theta) + \mathcal{L}_{\mathrm{ro}}(D_\theta) + \lambda_{\mathrm{con}}\,\mathcal{L}_{\mathrm{contact}}(D_\theta) + \lambda_{\mathrm{pen}}\,\mathcal{L}_{\mathrm{pen}}(D_\theta).$$

Reconstruction & Alignment

Diffusion Reconstruction

Ldiff : Reconstructs the clean hand–object motion from its noisy version.

Distance Map

Ldm : Preserves the spatial distances between the hands and the object.

Relative Orientation

Lro : Maintains consistent hand–object relative orientations.

Physical Plausibility

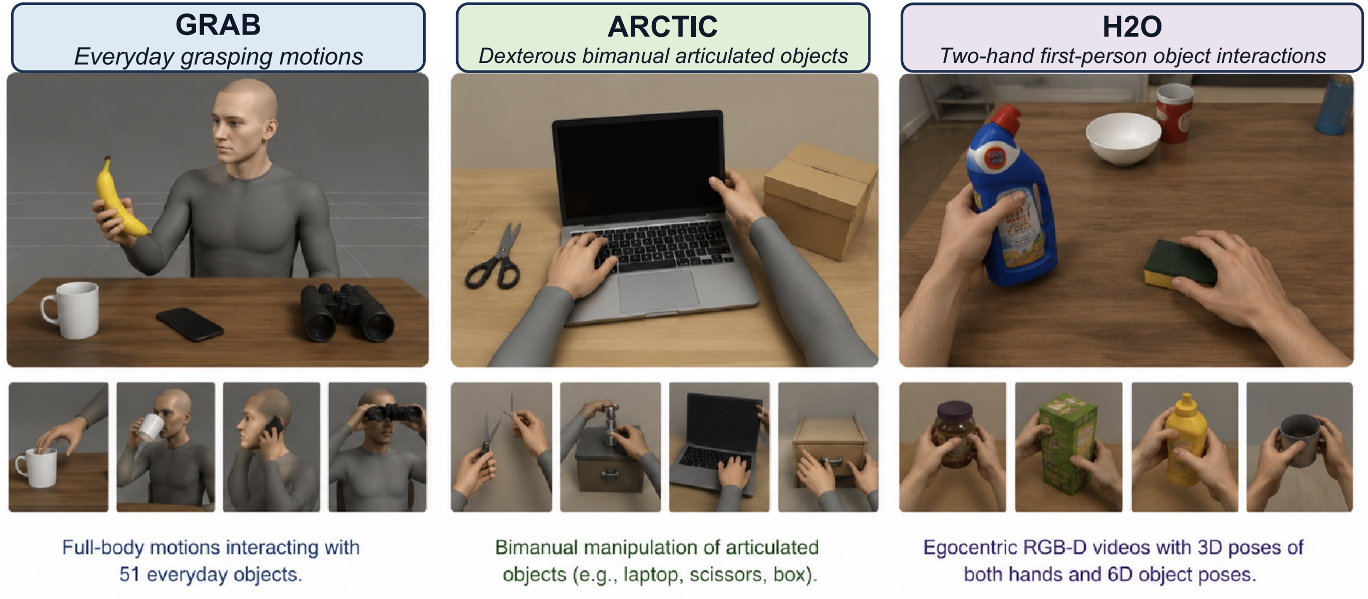
Contact

Lcontact : Encourages realistic hand–object contact during interaction.

Penetration

Lpen : Penalizes physically invalid hand–object interpenetration.

Data & Evaluation protocol



Metrics

- Motion quality: FID, Diversity
- Text alignment: R-Precision, MM-Dist

- Diversity under same prompt, MModality
- Geometry: Pen 1cm, Contact accuracy

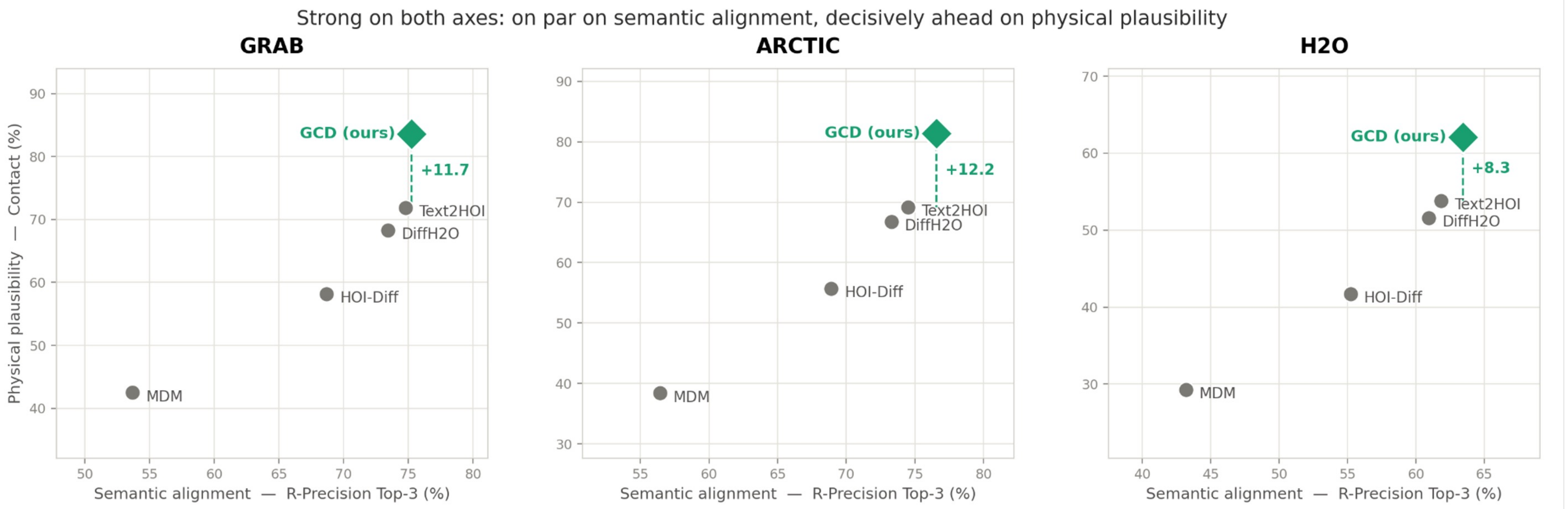
Quantitative Results

GRAB Dataset

Methods	FID ↓	Diversity →	R-Precision (Top 3) ↑	MM-Dist ↓	MModality ↑	Pen 1cm (%) ↓	Con (%) ↑
Real	0.01	17.03	88.76	1.95	-	**0.051**	**100.0**
HOI-Diff [13]	5.71 ± 0.46	11.25 ± 1.81	68.67 ± 0.02	6.10 ± 0.20	4.17 ± 0.18	0.76	58.2
MDM [20]	8.69 ± 1.06	10.69 ± 1.33	53.69 ± 0.01	8.97 ± 0.13	3.71 ± 0.18	3.76	42.5
DiffH2O [3]	3.28 ± 0.82	14.95 ± 2.05	73.42 ± 0.03	4.89 ± 0.24	5.42 ± 0.19	0.18	68.3
Text2HOI [1]	3.02 ± 0.76	**15.60 ± 2.18**	74.80 ± 0.02	4.71 ± 0.21	**5.76 ± 0.22**	0.13	71.9
GCD (Ours)	**2.93 ± 0.69**	15.93 ± 2.25	**75.25 ± 0.06**	**4.30 ± 0.19**	5.51 ± 0.26	**0.093**	**83.6**

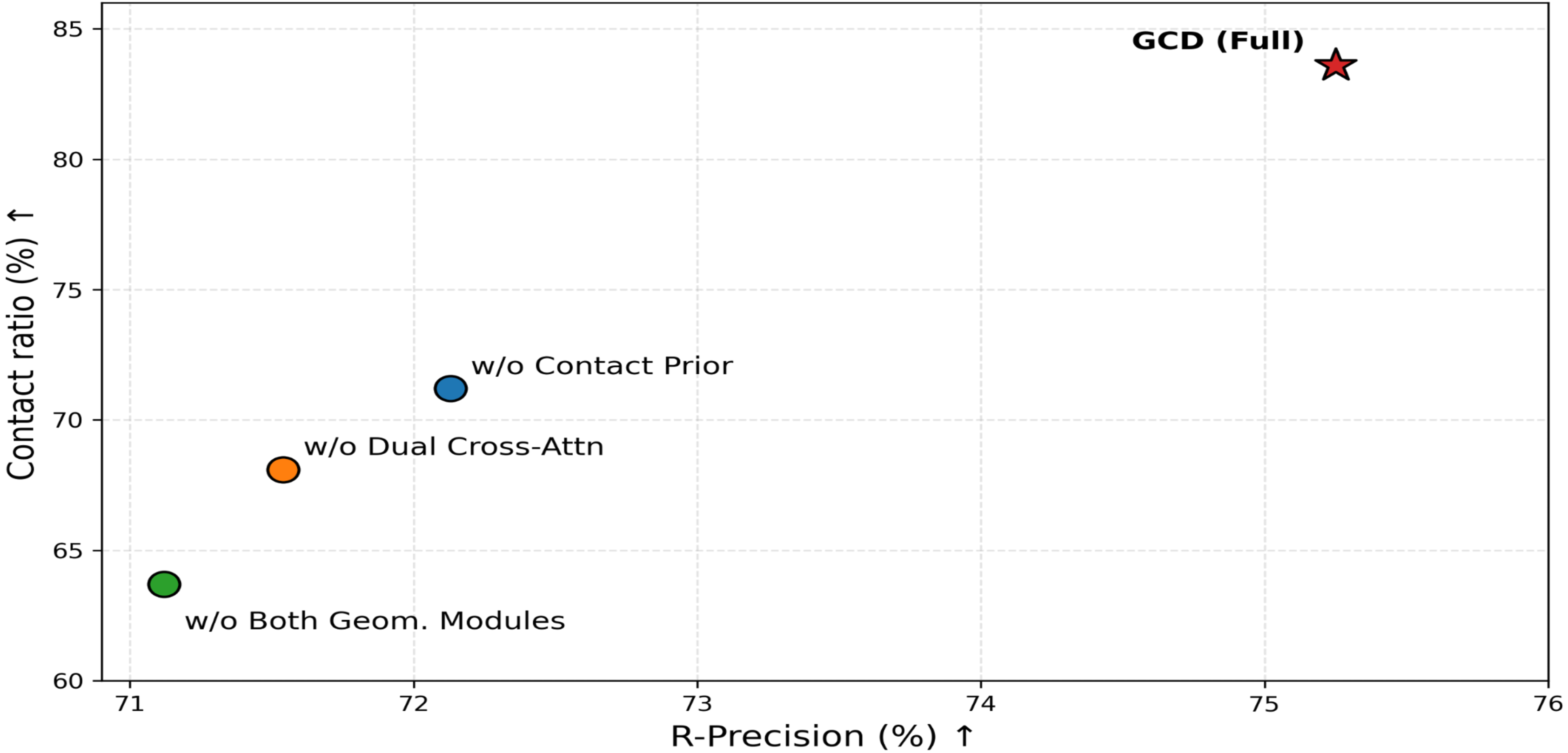
GCD remains competitive in motion–prompt alignment while improving physical realism through better contact and reduced penetration.

Results



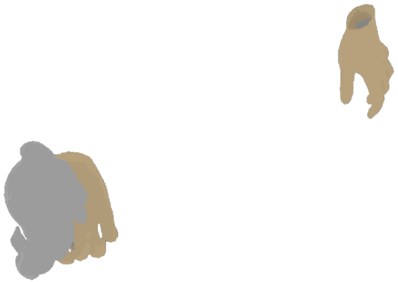
Strong on both axes: on par on semantic alignment, decisively ahead on physical plausibility
GRAB
ARCTIC
H2O
Physical plausibility — Contact (%)
Semantic alignment — R-Precision Top-3 (%)
GCD (ours)
+11.7
+12.2
+8.3
Text2HOI
DiffH2O
HOI-Diff
MDM

Ablations

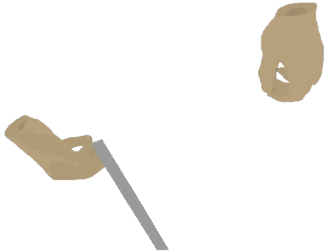

GCD (Full)
w/o Contact Prior
w/o Dual Cross-Attn
w/o Both Geom. Modules
Contact ratio (%) ↑
R-Precision (%) ↑
85
80
75
70
65
60
71
72
73
74
75
76

Qualitative RESULTS

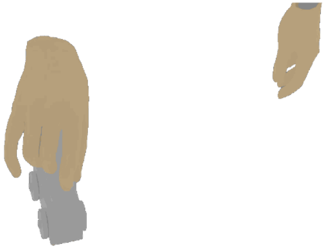
Inspect the Stanford Bunny with both hands.



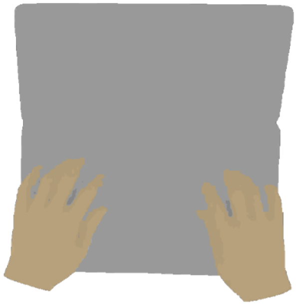
Play the flute with both hands.



Lift the train with both hands.

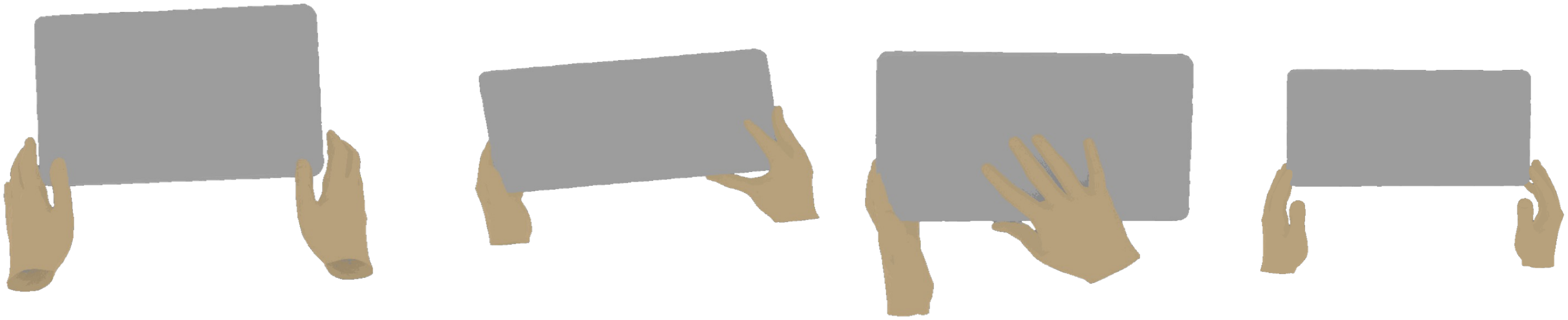


Type on the laptop with both hands.



RESULTS-Diversity

Prompt : Open laptop with both hands.



Diverse motions, physically plausible and geometrically consistent interactions.

Limitations and future directions

Contact is enforced as a geometric constraint, without explicitly modeling its temporal evolution.

↓

Model contact over time

Contact prediction degrades on substantially unseen object geometries.

↓

More transferable contact reasoning

Iterative diffusion sampling limits deployment in interactive applications such as animation, embodied agents, and VR.

↓

Toward faster interaction generation

Conclusion

✓ Geometry-aware dual-stream diffusion for text-driven hand–object interaction.

✓ Asymmetric conditioning: text guides object motion, while geometry guides hand motion.

✓ Geometry feedback during denoising improves contact quality and reduces interpenetration.

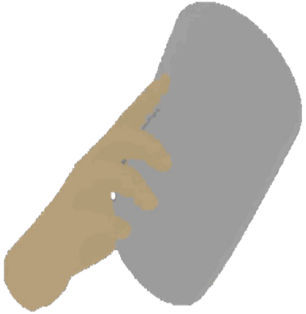
Thanks for your attention

Project

Unseen Objects Results

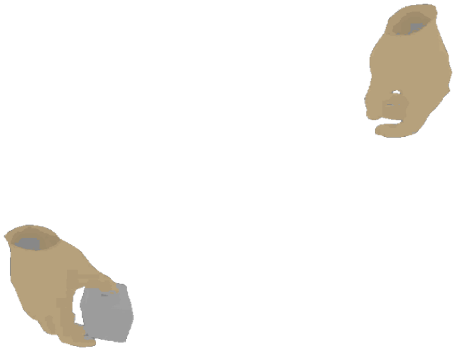
Transfer object from Grab to H2O motion generation:

Pour (milk Cup) with right hand.



Transfer object from H2O to Grab motion generation:

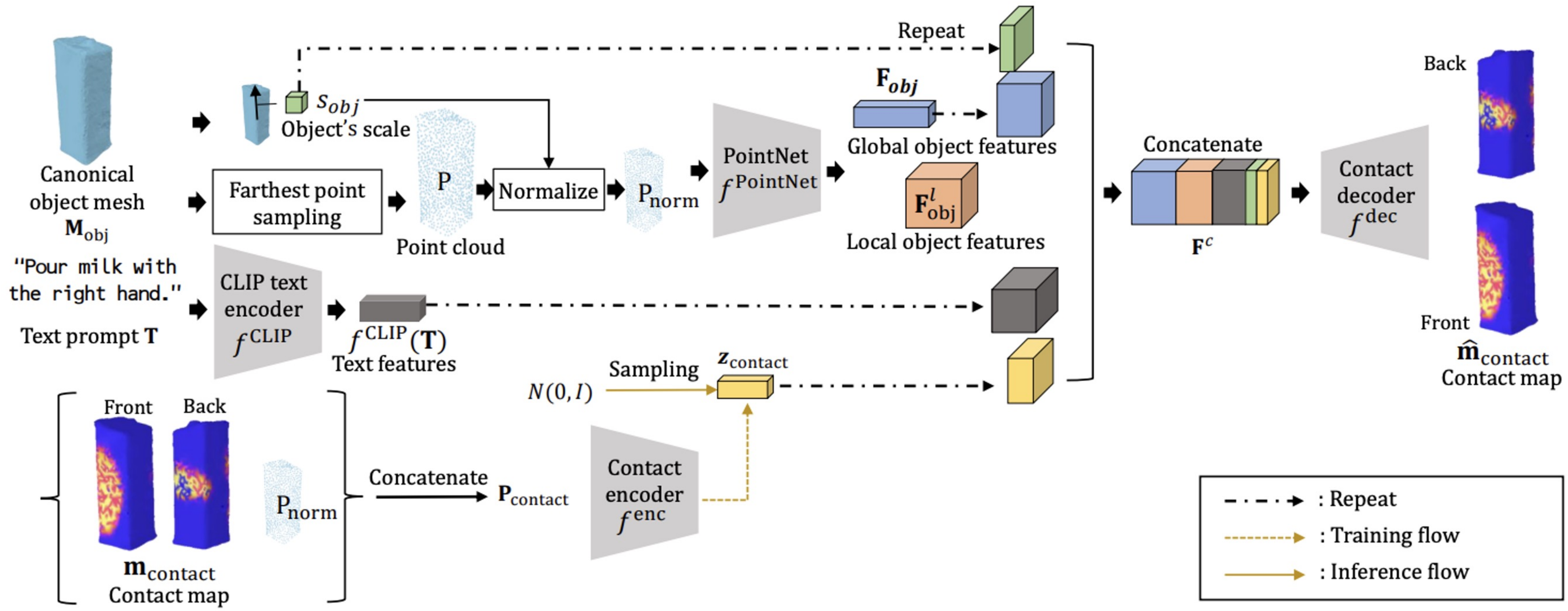
Drink (milk) with both hand.



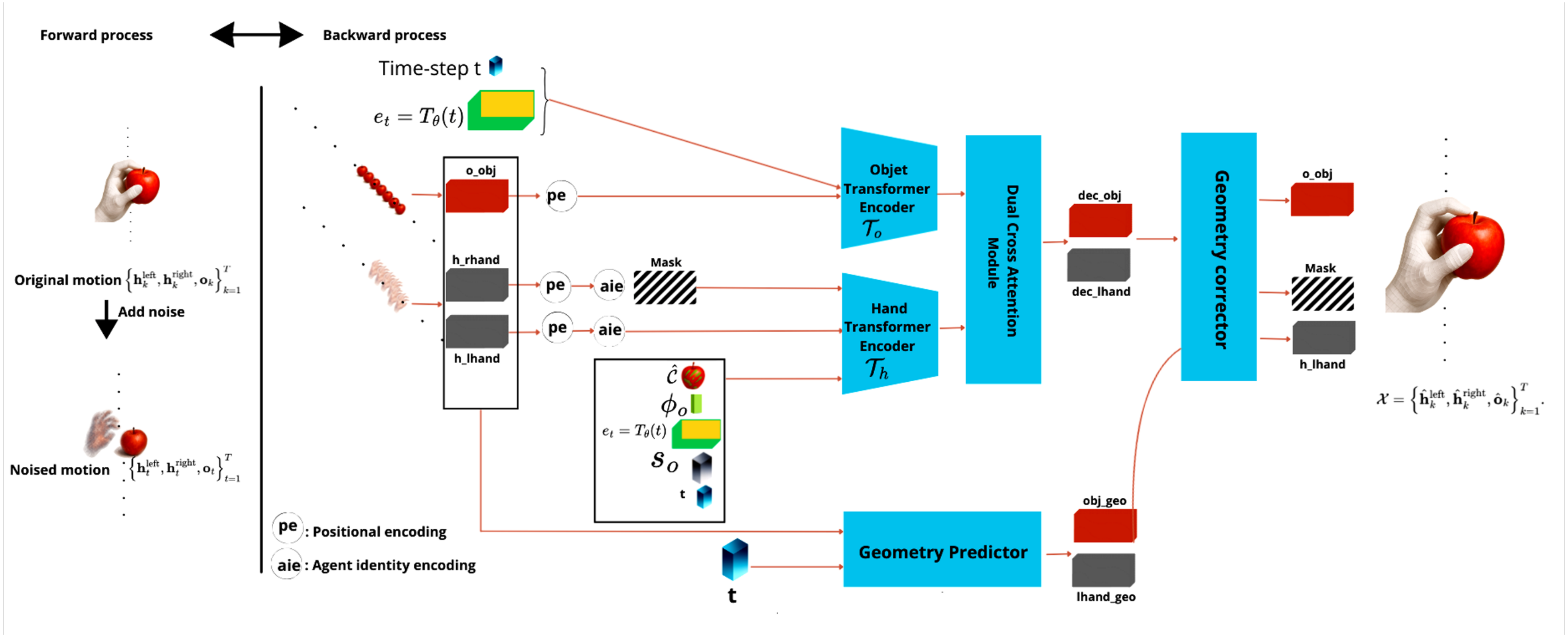
Transfer object from H2O to Grab motion generation:

Shake (chips) with both hand.

Contact Predictor Architecture



Motion Generator Architecture



References

- Cha, J., Kim, J., Yoon, J.S., Baek, S.: 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)
- Peng, X., Xie, Y., Wu, Z., Jampani, V., Sun, D., Jiang, H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2878–2888 (2025)
- Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
- Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
- Ho, J, Jain, A, Abbeel, P. Denoising diffusion probabilistic models. Advances in neural information processing systems 2020;33:6840–6851.
- Zhang, W, Dabral, R, Golyanik, V, Choutas, V, Alvarado, E, Beeler,T, et al. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. 2025, p. 27694–27705.