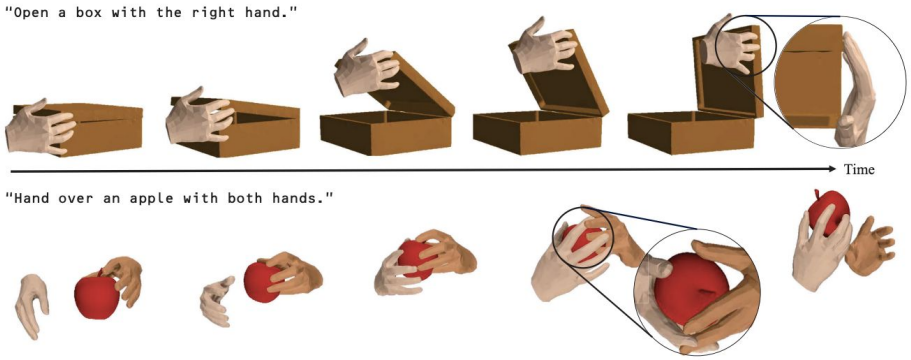
Text-guided 3D Motion Generation for Hand-Object Interaction

What is the Problem?

- **Lack of labeled data**: Current datasets do not cover the diverse hand-object interactions, both in terms of object categories and types of interaction.
- **Challenges in physical realism**: Generating physically accurate 3D interactions, ensuring correct hand-object contacts, and preventing artifacts (e.g., object penetration) are significant hurdles.
- **Semantic understanding**: Previous methods fail to capture the full semantic context of the interaction from text prompts.



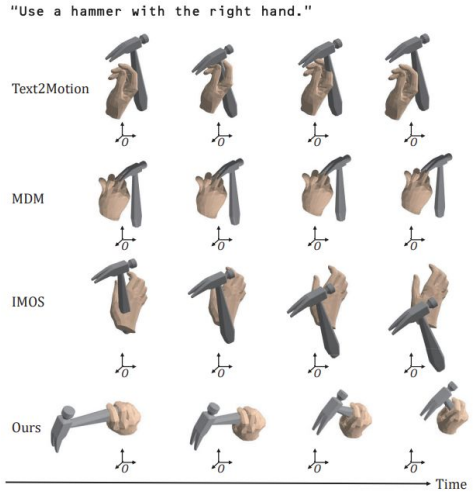
Earlier Work

- **Human body motion from text**: Previous methods focused on generating human motions from text (e.g., full-body or body-arm motion), without incorporating hand-object interaction.
- **Grasp-based methods**: Some approaches, like contact-based grasp synthesis, focus on static grasping but lack dynamic, realistic interactions or do not include motion generation.
- **Action label-based approaches**: Other methods use predefined action labels to generate hand or body motion but lack the flexibility to generalize to new, unseen objects.

B421048, Shubhank Nagar

Remaining Challenges

- **Data scarcity**: Existing datasets lack the necessary diversity to capture complex, real-world hand-object interactions, hindering generalization.
- **Motion generalization**: Generating interactions with unseen objects or new types of actions remains a problem, as many models are limited to predefined actions.
- **Artifact prevention**: Current methods struggle with ensuring the physical realism of interactions, particularly preventing penetration artifacts or maintaining consistent contact between hand and object.



Novel Solution

- **Decomposition of tasks**: The authors decompose interaction generation into two subtasks: generating a contact map between hand and object and generating motion based on this map.
- **Contact map generation**: A VAE-based network is used to predict the probability of contact points between hand and object surfaces, using a text prompt and object mesh.
- **Transformer-based diffusion model**: A diffusion-based model generates motion using the contact map, learning geometric and physical constraints to ensure plausible hand-object interactions.
- **Hand refinement module**: This module refines the generated motion to improve hand-object contact accuracy, reduce penetration artifacts, and enhance physical realism.
- **Compositional framework**: By breaking the task into modular components, the method can generalize better, handle unseen objects, and generate diverse interactions from text.