October 2025

# NovaFlow: Zero-Shot Manipulation via Actionable Flow from Generated Videos

Hongyu Li[1,2*], Lingfeng Sun[1*], Yafei Hu[1], Duy Ta[1], Jennifer Barry[1], George Konidaris[2] and Jiahui Fu[1]
[*]Equal contribution, [1]Robotics and AI Institute, [2]Brown University

## Abstract

**Enabling robots to execute novel manipulation tasks zero-shot is a central goal in robotics. Most existing methods assume in-distribution tasks or rely on fine-tuning with embodiment-matched data, limiting transfer across platforms. We present NovaFlow, an autonomous manipulation framework that converts a task description into an actionable plan for a target robot without any demonstrations. Given a task description, NovaFlow synthesizes a video using a video generation model and distills it into 3D actionable object flow using off-the-shelf perception modules. From the object flow, it computes relative poses for rigid objects and realizes them as robot actions via grasp proposals and trajectory optimization. For deformable objects, this flow serves as a tracking objective for model-based planning with a particle-based dynamics model. By decoupling task understanding from low-level control, NovaFlow naturally transfers across embodiments. We validate on rigid, articulated, and deformable object manipulation tasks using a table-top Franka arm and a Spot quadrupedal mobile robot, and achieve effective zero-shot execution without demonstrations or embodiment-specific training. Project website: https://novaflow.lhy.xyz/.**

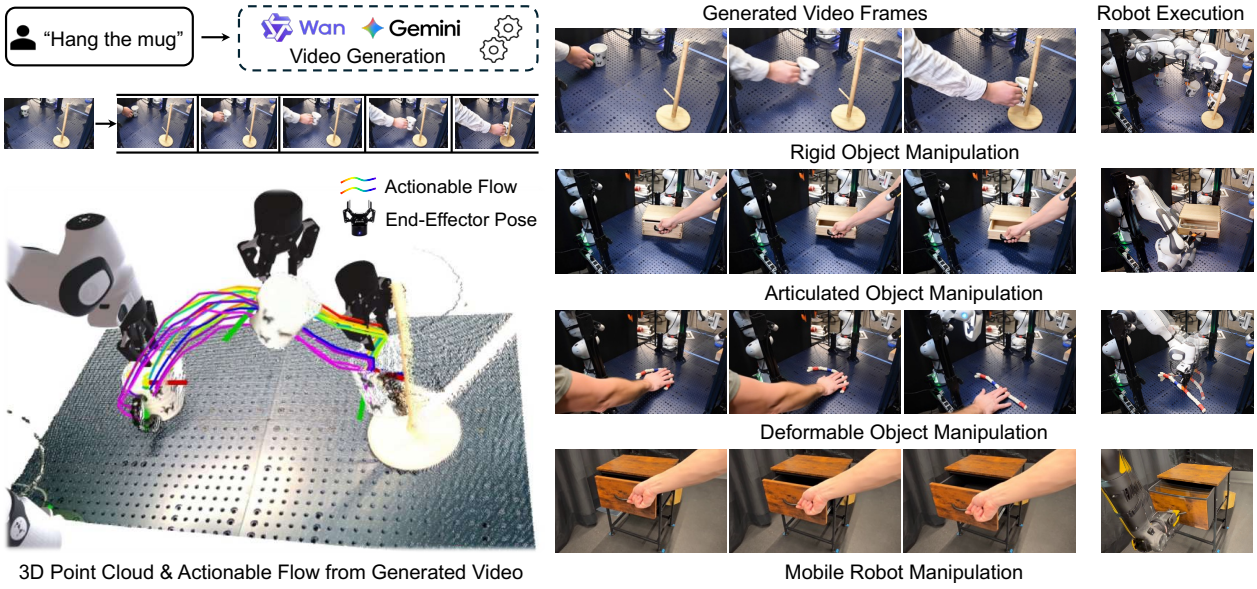


**Figure 1: NovaFlow manipulation framework.** A generated task-solving video is distilled into a **3D actionable object flow** aligned with the robot's observation. From this flow, reference end-effector trajectories are computed and tracked, enabling robots to manipulate rigid, articulated, and deformable objects across different embodiments *without demonstrations.*

*Corresponding author: Hongyu Li, George Konidaris ({hli230,gdk}@cs.brown.edu), Lingfeng Sun, and Jiahui Fu ({lsun, jfu}@rai-inst.com).*
*Hongyu Li and George Konidaris were supported by the Office of Naval Research (ONR) under REPRISM MURI N000142412603 and ONR grant N00014-22-1-2592. Partial funding was also provided by the Robotics and AI Institute.*

## 1. Introduction

A long-standing goal in robotics is to build generalist robots capable of performing a wide variety of manipulation tasks in unstructured environments without task-specific training. Many believe that Vision-Language-Action (VLA) models [1–4] can achieve this generalization, following the success of Large Language Models (LLMs) [5–7], Vision-Language Models (VLMs) [8, 9], and video generation models [10–13] that learn from vast, internet-scale datasets. However, directly applying this paradigm to robotics creates a significant data bottleneck. VLA models, for their end-to-end training, require vast quantities of robot-specific vision-language-action data that is difficult and expensive to collect, a stark contrast to the readily available web-scale data used for LLMs and VLMs.

An alternative path towards generalist robots lies in creating modular systems that decompose the problem into task understanding and robot control. These systems leverage powerful pretrained models [14, 15] and traditional robotic engineering methods like inverse kinematics (IK) [16] or model predictive control [17] to bypass large-scale robot data collection, a promising strategy for closing the data gap [18]. For instance, some approaches use large language or vision-language models to generate high-level plans, affordance maps, or semantic keypoints to guide the robot [19–22]. While these methods successfully offload semantic reasoning to large models, translating this understanding into physical actions remains an open problem. The control policy, for instance, relies on either predefined skill primitives (*e.g.*, opening a drawer) [15, 23] or learned skills from real-world demonstrations [14, 19, 20, 22, 24, 25]. This approach reintroduces the data bottleneck and limits generalizability and scalability.

To overcome these limitations, we propose **NovaFlow**, a novel framework that breaks the dependency on robot data to achieve autonomous manipulation. Our key insight is to *repurpose large-scale pretrained video generation models as a source of commonsense task understanding and implicit physical knowledge for deriving object motion.* We hypothesize that by training on internet-scale video data, these models have already captured a rich, generalizable understanding of task and object dynamics that can be leveraged for unseen objects, environments, and tasks. This separates our approach from prior work that relies on self-collected data to train smaller, specialized video models [26–29]. To translate this understanding from video to robot actions, we leverage *actionable 3D object flow*, a generalized atomic representation of object motion.

NovaFlow generates robot actions from a single visual observation and task description and consists of two components: a flow generator and a flow executor. The flow generator leverages large-scale video generation models to distill generalized knowledge of object motion into an *actionable 3D object flow*. This is achieved using a pipeline of pretrained perception modules for monocular depth estimation [30], 3D point tracking [31], and object grounding [32, 33]. The flow executor then translates this 3D flow into robot actions using IK and trajectory optimization, requiring no robot-specific data or task training. To handle diverse object types, the executor uses correspondence-based model-free tracking for rigid and articulated objects [34] and dynamic model-based planning with particle models for deformable objects [35, 36], using the flow as a tracking objective.

In summary, we present NovaFlow, an object-centric and embodiment-agnostic framework for autonomous manipulation that requires no task-specific tuning. We demonstrate its efficacy across both tabletop and mobile manipulator tasks involving rigid, articulated, and deformable objects. At its core, we introduce an actionable 3D object flow representation that is key to its generalizability and achieve state-of-the-art zero-shot performance on a range of real-world tasks, outperforming previous demonstration-free and data-dependent methods.

## 2. Related Work

We define an approach as zero-shot or demonstration-free if it does not require collecting any robot-specific data or task-specific training. While LLMs and VLMs have shown promising zero-shot capabilities, their embodied successors, VLA models, have yet to achieve the same level of generalization. Recent VLAs [37, 38] still rely on data collection to generalize on novel embodiments or camera views. This is due to the data bottleneck created by the end-to-end training nature of VLAs. To address this, we decouple the task understanding (Sec. 2.1) and robot control, bridged by an intermediate representation, the 3D object flow (Sec. 2.2).

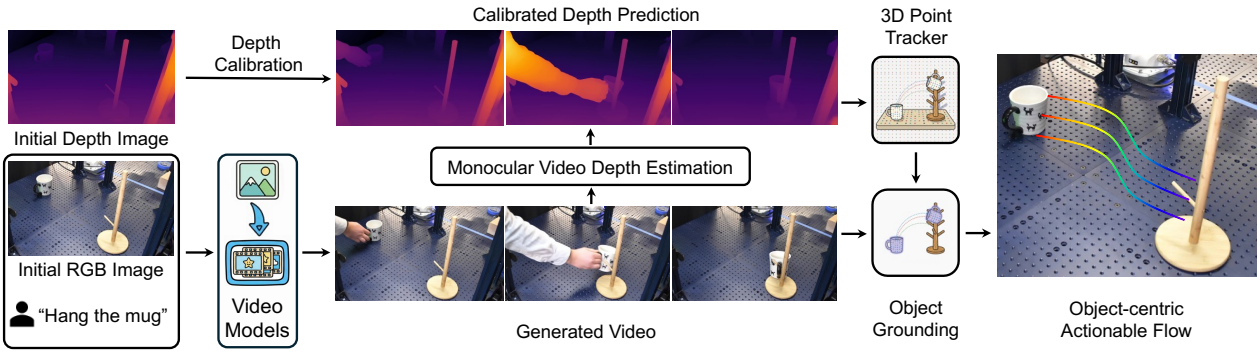


**Figure 2: Flow generator pipeline.** Given an initial image and a task prompt, a video model is used to generate a video of the plausible object motion. This video is then processed by pretrained perception modules to distill an actionable 3D object flow. This involves (1) lifting the 2D video to 3D using monocular depth estimation, (2) calibrating the estimated depth against the initial depth, (3) tracking the dense per-point motion using 3D point tracking, and (4) extracting the object-centric 3D flow via object grounding.

### 2.1. Video-based Manipulation

Prior work has utilized video generation models for manipulation. Video can be a generalized representation of motion, which serves as a visual instruction for robots to execute tasks. Some work trains an inverse dynamics model [26, 39] or a policy [28, 40, 41] to convert the generated video into robot actions. Other work tracks the 6D pose of the end-effector [27] or the object [16]. While promising, these approaches require extensive robot-specific data to train a domain-specific model tailored to a particular embodiment, environment, or task [26–28, 39–41].

A key limitation of many video-based manipulation methods is their reliance on embodiment-dependent action generation, which hinders cross-embodiment generalization. To address this, object-centric approaches have been proposed. For example, a concurrent work [16] extracts 6D poses from the generated video for demonstration-free manipulation, which is object-centric and generalizes across embodiments. However, it is model-based and relies on a rigid-body assumption, limiting its applicability to a broader class of objects. To achieve greater object generalization, a shift towards model-free representations is essential, which then motivates the adoption of flow-based approaches.

### 2.2. Flow-based Manipulation

Flow describes object motion by tracking the displacement of 2D pixels or 3D points between video frames. This offers a more generalizable representation of object dynamics compared to 6D pose, as it is inherently model-free and makes no assumptions about object rigidity. Recent work has shown success in using flow for manipulation [24, 29, 42–45]. However, these methods require robot data or task-specific training for either the flow generator or the executor [24, 29, 42–46]. To achieve greater generalization for zero-shot manipulation, Chen et al. [47] and Zhi et al. [48] train a flow generator on a collection of large-scale human egocentric datasets. While making a great step towards generalization, we empirically find that the generalizability of this approach [47] (understanding of in-the-wild object motion) is still not as good as the commonsense motion knowledge from pretrained video models.

## 3. NovaFlow

NovaFlow enables robots to autonomously solve a wide variety of manipulation tasks by leveraging pretrained video generation models, thus eliminating the need for demonstrations or task-specific tuning. Since raw video pixels cannot be directly used by a robot's controller or model-based planner, NovaFlow handles this challenge by distilling the video's implicit commonsense knowledge of motion into a more actionable, intermediate representation: 3D object flow. The proposed pipeline consists of two core components: a *flow generator* (Fig. 2) that extracts the actionable 3D object flow from the generated video, and a *flow executor* (Fig. 3) that translates this flow into robot actions. The entire pipeline is demonstration-free and embodiment-agnostic, requiring no robot-specific data or training before task execution.

### 3.1. Flow Generator

The primary objective of the flow generator is to translate a high-level task description into a struc-

tured, actionable flow for the robot. The standard input to the generator is the task description, which involves an initial RGB-D image pair $\{\mathbf{I}, \mathbf{D}\}$ captured from the robot's perspective (with known camera intrinsics) and a natural language instruction, $\mathbf{l}$, describing the desired task. For tasks requiring greater precision, an optional goal image, $\mathbf{I}_g$, can also be provided, either specified by the user [45, 49, 50] or generated by an image editing model with the standard input [51]. Based on the given input, the generator's goal is to produce a 3D object flow across $T$ frames for $M$ object keypoints, $\mathcal{F} \in \mathbb{R}^{T\times M\times 3}$.

The flow generator synthesizes a video using the task description and then distills an *actionable 3D object flow* from the generated video with a pipeline of pre-trained perception modules. The full process involves five main steps: (1) generating the video from the initial image and text prompt, (2) lifting the 2D video to 3D, (3) calibrating the estimated depth, (4) tracking dense per-point 3D motion, and (5) extracting the final object-centric 3D flow.

#### 3.1.1. Video Generation

Given the initial image $\mathbf{I}$ and language prompt $\mathbf{l}$, a video generation model produces a video $\hat{\mathbf{V}} = \{\hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2, \dots, \hat{\mathbf{I}}_T\}$ of $T$ frames, known as image-to-video (I2V) generation. If a goal image $\mathbf{I}_g$ is provided, we use first-last-frame-to-video (FLF2V) generation instead.

#### 3.1.2. Monocular Depth Estimation

To obtain 3D motion information, we lift the generated 2D video into 3D space. We apply a monocular video depth estimation model to $\hat{\mathbf{V}}$, which processes the video frame-by-frame to yield a sequence of estimated metric depth maps $\hat{\mathbf{D}} = \{\hat{\mathbf{D}}_1, \hat{\mathbf{D}}_2, \dots, \hat{\mathbf{D}}_T\}$.

#### 3.1.3. Depth Calibration

The depth maps $\hat{\mathbf{D}}$ obtained in the last step have a key limitation: the monocular depth estimation process is inherently ill-posed and often creates metric outputs with systematic scaling errors, especially on generated videos. This can hinder manipulation tasks that require accurate spatial alignment. To correct for this, we calibrate the entire estimated depth sequence $\hat{\mathbf{D}}$ by anchoring it to the initial ground-truth depth map. This calibration leverages the observation that estimated depth, while globally inaccurate, is often locally consistent. We compute a scaling factor between the median depth of the first estimated frame $\hat{\mathbf{D}}_1$ and the initial ground-truth depth map $\mathbf{D}$. While other methods exist, e.g., fitting an affine transformation [16], we find this median scaling factor method to be more stable.

#### 3.1.4. 3D Point Tracking

With the generated video and the calibrated depth, we extract dense per-point 3D motion. We employ a 3D point tracking model, which takes the camera intrinsics, video $\hat{\mathbf{V}}$, the calibrated depth $\hat{\mathbf{D}}$, and a set of query points $\mathcal{Q} = \{\mathbf{q}_1, \dots, \mathbf{q}_M\}$ evenly sampled on the first frame as input. The model outputs a set of 3D trajectories $\mathcal{P} = \{\mathbf{p}_1^t, \dots, \mathbf{p}_M^t\}_{t=1}^T$, where $\mathbf{p}_i^t$ is the 3D position of the $i$-th query point at timestep $t$.

#### 3.1.5. Object Grounding

The dense 3D trajectories $\mathcal{P}$ capture the motion of the entire scene. To derive an actionable plan, we must now ground this motion by isolating only the trajectories belonging to the target objects. We achieve this by employing a pipeline that combines an open-vocabulary object detector with a video segmentation model, which produces a sequence of masks, $\mathcal{M} = \{\mathbf{m}_1, \dots, \mathbf{m}_T\}$, that segment the object across the entire video. Lastly, by applying these masks, we filter the dense trajectories $\mathcal{P}$ to distill the actionable 3D object flow, $\mathcal{F} = \{\mathbf{f}_i^t \mid i = 1, \dots, K;\ t = 1, \dots, T\}$. This final output represents the $K$ keypoints that remain consistently tracked on the object's surface.

We conclude object grounding with a rejection sampling step to filter out hallucinations, such as generative artifacts and implausible motions, that may be unavoidably introduced by the video generation model (Fig. 4). Here, we use a VLM to validate and select the most plausible generated flow. Specifically, we generate $N$ video candidates simultaneously and obtain $N$ corresponding object flow images by back-projecting object flow $\mathcal{F}$ to the first frame of each video. We then mark each flow image using its ID and pass it into a VLM along with its task description to select the most plausible one. We empirically find that rejecting the flow image is more effective than rejecting the concatenated raw video images [16], since the flow images, explicitly visualizing motion, are a more grounded and direct input for VLMs to reason and understand.

### 3.2. Flow Executor

The flow executor is responsible for translating the abstract 3D object flow $\mathcal{F}$ as planned trajectories into a sequence of executable robot actions $\mathcal{A}$ =

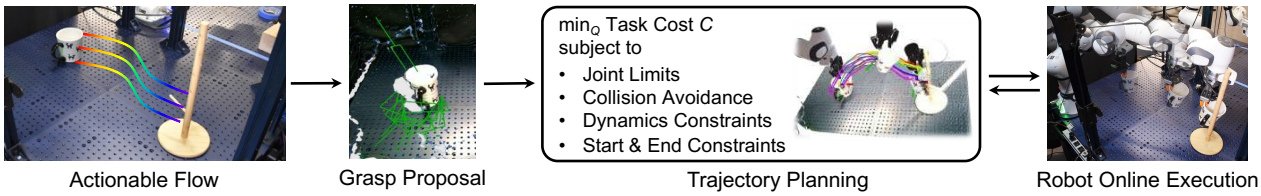


**Figure 3: Flow executor pipeline.** The initial end-effector pose is determined from grasp proposal candidates. Robot trajectories are then planned based on the actionable flow considering costs and constraints, and subsequently tracked by the robots.

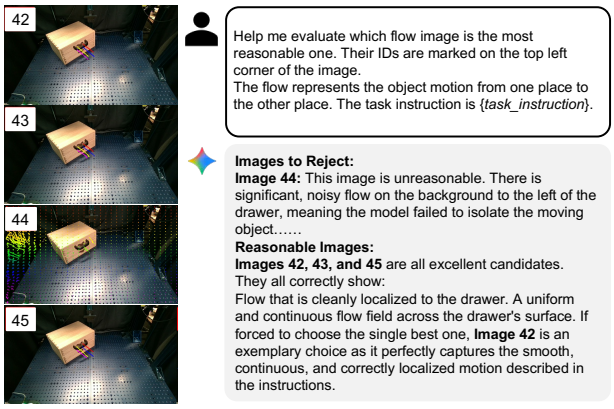


**Figure 4: Rejection sampling for flow generator.** We generate multiple video candidates in parallel and create the object flow image for each by back-projecting its object flow, $\mathcal{F}$, onto the initial frame. A VLM (in our case, Google Gemini) evaluates all the flow images to select the most plausible video candidate.

$\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_T\}$ over $T$ timesteps. The input actionable flow serves as an intermediate representation that describes the desired trajectories of $K$ points on the target object over $T$ timesteps, carrying the high-level task understanding from pre-trained video models. Here, we present an open-loop planner that can be extended to a closed-loop tracking system by incorporating live object trackers. The current executor pipeline can handle two main classes of objects: rigid (including articulated objects, treated as part-wise rigid) and deformable.

### 3.2.1. Rigid Object Manipulation

For rigid body manipulation, the 3D flow of the keypoints, $\mathcal{F}$, can be used to estimate the rigid transforms, $(\mathbf{R}, \mathbf{t})$, of the object across frames in a model-free manner. In cases where the object is firmly grasped and moves rigidly with the end-effector (*e.g.*, no slippage), a common assumption in prior work [16, 25, 44, 45], the end-effector pose can be calculated from the object pose. The object-specific firm grasps are selected from the object point cloud using a grasp proposal model, as shown in Fig. 3. At each timestep $t$, we find the rigid transformation $(\mathbf{R}^t, \mathbf{t}^t)$ that aligns the initial keypoints $\{\mathbf{f}_i^1\}_{i=1}^K$ to the current one $\{\mathbf{f}_i^t\}_{i=1}^K$. This is solved using the Kabsch algorithm [34], which finds the optimal rotation $\mathbf{R}^t$ that minimizes the sum of squared errors:

$$\mathbf{R}^t = \underset{\mathbf{R} \in SO(3)}{\operatorname{argmin}} \sum_{i=1}^{K} \|\mathbf{R}(\mathbf{f}_i^1 - \mathbf{c}^1) - (\mathbf{f}_i^t - \mathbf{c}^t)\|^2, \quad (1)$$

where $\mathbf{c}^1$ and $\mathbf{c}^t$ are the masked point cloud centroids at the first and current timesteps, respectively. This optimization can be solved efficiently using Singular Value Decomposition (SVD). Once the rotation is found, the translation is computed as $\mathbf{t}^t = \mathbf{c}^t - \mathbf{R}^t\mathbf{c}^1$. The object pose at timestep $t$ can be represented as a homogeneous transformation matrix $\mathbf{T}_{obj}^t \in SE(3)$, constructed from $\mathbf{R}^t$ and $\mathbf{t}^t$. The resulting sequence of 6D object poses is converted into an end-effector trajectory by applying a grasp transformation, $\mathbf{T}_{grasp}$, obtained from a grasping network [52, 53]. The target end-effector pose at each timestep is then:

$$\mathbf{T}_{ee}^t = \mathbf{T}_{obj}^t \cdot \mathbf{T}_{grasp}. \quad (2)$$

This Cartesian pose is converted to joint commands via trajectory optimization for execution by the robot's controller.

### 3.2.2. Deformable Object Manipulation

Unlike rigid objects, deformable objects have complex dynamics that cannot be described by a simple rigid transformation. NovaFlow can be naturally extended to handle deformable objects, with the 3D object flow $\mathcal{F}$ serving as a dense tracking objective for model-based planning. Specifically, we employ a particle-based dynamics model $f_\theta$ to predict the object's future state, where $\theta$ represents the learnable parameters of the model. The state of the object at time $t$ is represented by a set of $N_p$ particles $\mathcal{S}_t = \{\mathbf{s}_i^t\}_{i=1}^{N_p}$. The dynamics model predicts the next state based on the current state and a robot action $\mathbf{a}_t$: $\mathcal{S}_{t+1} = f_\theta(\mathcal{S}_t, \mathbf{a}_t)$.

Conventional methods for deformable manipulation often define a cost function using a correspondence-free metric, like the Chamfer distance, to a single goal state $\mathcal{S}_{goal}$ [35, 36, 54–56]. Our actionable 3D object flow $\mathcal{F} = \{\mathcal{F}^t\}_{t=1}^{T}$, where $\mathcal{F}^t = \{\mathbf{f}_i^t\}_{i=1}^{N_p}$, allows us to define a cost function based on the sum of squared Euclidean distances, leveraging the explicit point-wise correspondences from the flow:

$$C(\mathcal{S}_t, \mathcal{F}^t) = \sum_{i=1}^{N_p} \|\mathbf{s}_i^t - \mathbf{f}_i^t\|^2. \tag{3}$$

This formulation has two potential advantages. First, using point correspondences may create a better-conditioned optimization landscape, as correspondence-free metrics can be susceptible to local minima. Second, tracking a dense flow provides intermediate targets along a desired motion path, rather than relying on only a final goal configuration.

We then frame the control problem as a Model Predictive Control (MPC) task. At each timestep $t$, we solve for an optimal sequence of actions $\mathbf{A}_t^* = \{\mathbf{a}_t^*, \dots, \mathbf{a}_{t+H-1}^*\}$ over a planning horizon $H$ by minimizing the cumulative cost:

$$\mathbf{A}_t^* = \underset{\mathbf{A}_t}{\operatorname{argmin}} \sum_{j=t}^{j+H-1} C(\mathcal{S}_j, \mathcal{F}^j), \tag{4}$$

subject to the *dynamics constraints* $\mathcal{S}_{j+1} = f_\theta(\mathcal{S}_j, \mathbf{a}_j)$. We then execute the first action $\mathbf{a}_t^*$ and repeat the optimization at the next timestep.

**Trajectory optimization.** To enable smooth and collision-free motion, we additionally incorporate trajectory optimization to refine the sequence of actions. We formulate the trajectory generation as a non-linear least-squares problem. The goal is to find an optimal sequence of joint configurations $Q = \{q_0, q_1, \dots, q_{T-1}\}$ that minimizes a sum-of-squares objective function. The trajectory is initialized by linearly interpolating between start and end configurations, $q_{\text{start,IK}}$ and $q_{\text{end,IK}}$, which are precalculated using an IK solver using the end-effector pose. The optimal trajectory $Q^*$ is found by solving the constrained non-linear optimization problem:

$$\begin{aligned} \min_{Q} \quad & w_s C_{\text{smooth}} + w_r C_{\text{rest}}, \quad \text{subject to} \\ & q_0 = q_{\text{start,IK}} \quad \text{and} \quad q_{T-1} = q_{\text{end,IK}}, \\ & q_{\min} \le q_t \le q_{\max}, \quad \forall t \in \{0, \dots, T-1\}, \\ & d_s(q_t, q_{t+1}, O_j) \ge \epsilon_{\text{safe}}, \quad \forall t, \forall O_j \in \text{Obstacles}. \end{aligned} \tag{5}$$

In this formulation, the objective function seeks to minimize a weighted sum of the motion smoothness cost ($C_{\text{smooth}}$) and the rest pose regularization cost ($C_{\text{rest}}$). Constraints in the optimization include: (1) *start and end constraints*, meaning the trajectory's start and end configurations ($q_0$ and $q_{T-1}$) must exactly match the predefined goals ($q_{\text{start, IK}}$ and $q_{\text{end, IK}}$); (2) *collision avoidance*, by enforcing the signed distance, $d_s$, between the robot and any obstacle to remain greater than a safety margin, $\epsilon_{\text{safe}}$, at all times; (3) *joint limits*, ensuring the robot's physical joint position and velocity limits throughout the entire motion. We treat these constraints as cost terms and use the Levenberg-Marquardt solver to solve the non-linear least-squares problem.

## 4. Experiments

We aim to demonstrate the generalizability of NovaFlow across different object types and embodiments and to show the importance of each component in our framework. We evaluate the framework's ability to execute a broad range of manipulation tasks involving rigid, articulated, and deformable objects across embodiments without requiring task-specific demonstrations or additional fine-tuning.

### 4.1. Implementation Details

We implement NovaFlow with modular, swappable components. For video generation, we use the open-source model Wan [13], which produces 41 frames per task (16 FPS, 1280×720). We estimate depth with MegaSaM [30] using calibrated intrinsics, track 3D points with TAPIP3D [31], and ground objects via Grounded-SAM2 [58] (Grounding DINO [32] + SAM2 [33]). We use a trained PhysTwin [35] model to predict particle dynamics for deformable objects. All modules are drop-in replaceable with newer models, improving speed and robustness, which is another benefit for our modular framework.

### 4.2. Real-World Experiments and Evaluation Tasks

We evaluate NovaFlow on a Franka arm with a Robotiq-85 gripper for table-top manipulation and a Spot quadruped for mobile robot manipulation. For rigid and articulated objects, we use a single RealSense D455 depth camera as input. For deformable objects, we use three synchronized cameras (as required by PhysTwin [35]), though a single-view setup is also possible [56].

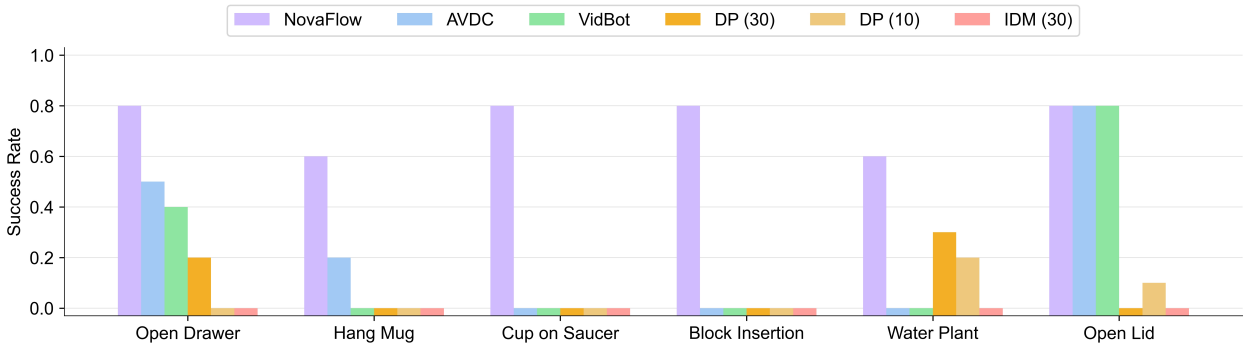


**Figure 5: Experiment results.** We compare against Diffusion Policy (DP) [57] trained using 10 and 30 demonstrations, inverse dynamics model (IDM) from UniPi [26], AVDC [29], and VidBot [47] in real-world tabletop manipulation tasks.

We categorize our tasks by the object type involved as rigid (**R**), articulated (**A**), and deformable (**D**):

- **Hanging a mug (R)**: hang a mug on a wooden rack, requiring accurate relative pose placement for the handle to pass through the wooden stick on the rack.
- **Inserting a block (R)**: insert a yellow block into a hole in a board, a task similar to peg-in-hole that requires accurate insertion skills.
- **Placing a cup on a saucer (R)**: place a cup on a saucer, a task demanding accurate placement skills.
- **Watering a plant (R)**: pour water from a green cup into a plant pot, requiring language understanding and manipulation skills.
- **Opening a drawer (A)**: open a drawer, requiring a precise understanding of its articulation.
- **Straightening a rope (D)**: straighten a curved rope, which requires understanding the dynamics of a deformable object.

During evaluation, we randomize the object placement after each trial. We report the quantitative and qualitative results in Fig. 5 and Fig. 6.

### 4.3. Comparison with Baselines

We compare NovaFlow against two groups of baselines. (+) denotes methods requiring external training data, while (*) denotes baselines adapted to fit our pipeline.

**Demo-free, zero-shot baselines (similar to ours):**

- **AVDC [29] (*):** Extracts object-centric motion using optical flow. We adapt it to our pipeline by applying it directly to generated videos.
- **VidBot [47]:** Learns flow from large-scale human interaction datasets to model affordances.

**Data-dependent baselines (require demonstrations):**

- **Diffusion Policy (DP) [57] (+):** Diffusion policy serves as an imitation policy baseline trained under very few demos for a single task. We train with 10 and 30 demonstrations per task, using the same single-view camera RGB input as our approach.
- **Inverse Dynamics Model (IDM) [26] (+):** IDM was originally designed to train together with a fine-tuned video generation model using in-domain demonstrations. Since video fine-tuning is outside our scope, we trained the IDM model with the 30 demonstrations previously used in DP training to convert generated robot task-solving videos (from Wan2.1) into robot actions.

NovaFlow achieves the highest success rates across tasks among zero-shot methods and also surpasses data-dependent baselines trained with 10–30 demonstrations, as shown in Fig. 5 (with 10 trials for each task). **AVDC(*)** performs competitively on affordance-like tasks but struggles with precise, long-horizon placements. In our setup, it distills motion from 2D optical flow, lacking 3D awareness and long-term coherence under occlusion. These limitations, as also noted in the AVDC paper, cause the method to struggle with tasks requiring accurate placement and rotation-heavy motions. **VidBot** excels on affordance-centric, articulated interactions (e.g., "open drawer") but fails when tasks require object–object relations and precise relative pose placement. This matches our diagnosis that its training emphasizes object–affordance understanding rather

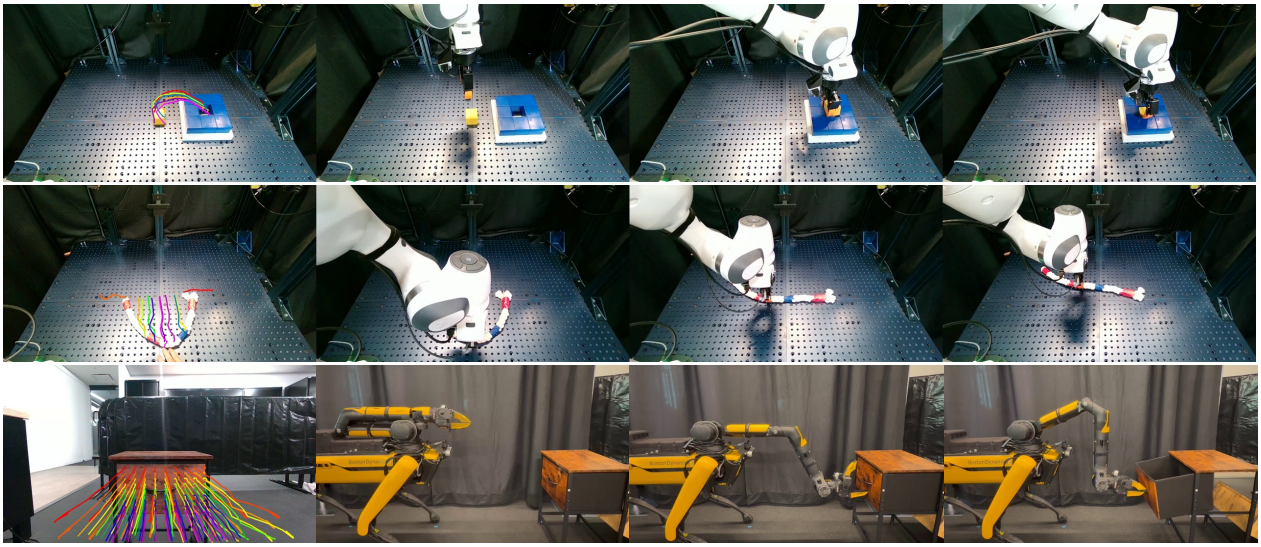

**Figure 6: Real-world manipulation experiments.** NovaFlow is versatile and supports cross-embodiment manipulation, which we use to manipulate rigid, deformable, and articulated objects using tabletop and mobile manipulator.

than modeling multi-object constraints. For **DP(+)**, despite per-task training (an easier setting that bypasses language understanding), it still shows poor generalization from a few examples, especially because our evaluations are randomly sampled and not drawn from the training distribution. The main issue for **IDM(+)** is the domain shift between its training and test data. The inverse dynamics model learns from real-world robot demos, yet it must interpret generated videos whose motion is not always kinematically perfect or consistent. Consequently, the generated videos are out-of-distribution, causing the model to fail even if the video's high-level action seems semantically reasonable.

Overall, methods that (i) lack an actionable *3D* representation (AVDC, VidBot) or (ii) rely on small, task-specific robot datasets (DP, IDM) fail to provide zero-shot autonomous task-solving. Distilling a dense, actionable 3D object flow and decoupling understanding from control is critical for zero-shot generalization.

### 4.4. Ablation Studies and Failure Analysis

Here we discuss some of the design choices in NovaFlow's submodules. For video generation, we compare the current Wan 2.1 model to the closed-source model Veo [59], which produces 8 s clips (24 FPS). Prompt extension is utilized for better controllability. For precise placement tasks (e.g., mug on rack and block insertion), we optionally condition on a goal image (FLF2V) instead of I2V.

We analyze the failure cases of NovaFlow in Fig. 7,

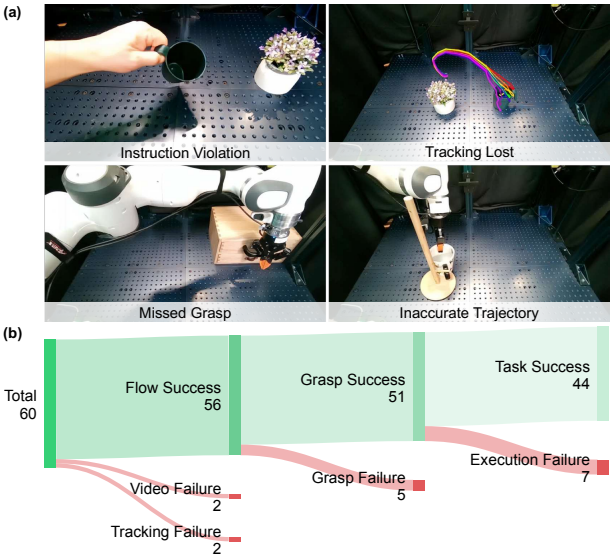


**Figure 7: Failure analysis.** (a) Examples of video, tracking, grasp, and execution failures. (b) Failure cause distribution.

identifying four primary failure modes. **Video failure** occurs when the generative model produces content that is not physically plausible, lacks 3D consistency, or violates the user's instructions; our rejection sampling with a VLM mitigates but does not eliminate this. **Tracking failure** results from inaccuracy in 3D point tracking, often caused by textureless surfaces, heavy occlusions, or accumulated inconsistencies inherited from the video model. **Grasp failure** happens when the robot fails to secure the object correctly (e.g., bad approach, missed grasp, and slip). Finally, **execution failure** encompasses errors during trajectory execution, such as

**Table 1:** Effect of goal image on block insertion task.

| Condition | Video Success | Task Success | Time (s) |
|---|---|---|---|
| w/ Goal Image (Wan2.1) | 46% | 80% | 612 |
| w/o Goal Image (Wan2.1) | 15% | 40% | 612 |
| w/o Goal Image (Veo) | 75% | 80% | 20 |

**Table 2:** Runtime analysis. Time is measured in seconds.

| | MegaSaM | TAPIP3D | SAM2 | **Total (Veo)** | Total (Wan) |
|---|---|---|---|---|---|
| **Time** | 100 | 5 | 8 | **133** | 725 |

collisions, joint limits, or an inability to follow the planned path accurately. Our analysis reveals that most failures occur in the *last mile*: grasp and execution are the most frequent, suggesting that while the upstream flow estimation is relatively robust, physical interaction remains the bottleneck. This is similar to the sim-to-real gap in simulation-based training. To address these limitations, future work could focus on integrating a closed-loop feedback system to enable dynamic replanning and refine the generated flow in response to observations.

We investigate the effect of a goal image on the block insertion task requiring millimeter-level precision (Tab. 1). We use two metrics: Video Success Rate, the percentage of generated videos with a valid actionable flow, and Task Success Rate, the execution success of a flow selected by a VLM after rejection sampling. For each trial, we synthesize eight videos, from which the VLM selects the best for execution. Our results show that omitting the goal image significantly impairs the performance of the open-source Wan2.1 model. While VLM rejection sampling improves the final success rate, the drop remains substantial. In contrast, the closed-source Veo model proves more robust, outperforming Wan2.1 and achieving a high task success rate even without a goal image.

### 4.5. Runtime Analysis

We deploy NovaFlow on a single NVIDIA H100 GPU, and a complete flow generation takes around 2 minutes end-to-end (Veo). We report per-module timings in Tab. 2 to guide replacements and optimization. The dominant time-consuming modules are the video generation and 3D lifting modules. For video generation, closed-source models are usually much faster in time but more expensive in cost.

## 5. Conclusion

We introduced NovaFlow, a demonstration-free framework for autonomous manipulation that translates natural language commands into robot actions by leveraging the commonsense knowledge embedded in large-scale video generation models. Our key insight is to distill generated task-solving videos into an actionable 3D object flow, an intermediate representation that decouples high-level task understanding from low-level robot control. This modular design enables NovaFlow to handle rigid, articulated, and deformable objects across different robot embodiments without requiring any task-specific training or demonstrations. Our real-world experiments show that NovaFlow not only outperforms other zero-shot methods but also surpasses imitation learning policies trained on dozens of demonstrations.

Despite its success, our failure analysis reveals that the primary bottleneck is the physical execution phase, particularly in grasping and handling unexpected dynamics. This highlights a gap between the open-loop plan generated from video and the complexity of real-world interaction. A promising direction for future work is to develop a closed-loop system where real-time feedback from the environment is used to refine or replan the generated flow, making the system more adaptive and robust to unforeseen challenges.

## References

[1] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi *et al.*, "OpenVLA: An Open-Source Vision-Language-Action Model," in *Conference on Robot Learning (CoRL)*, Jan. 2025.

[2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog *et al.*, "RT-1: Robotics Transformer for Real-World Control at Scale," Dec. 2022, arXiv:2212.06817 [cs].

[3] D. Ghosh, H. R. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu *et al.*, "Octo: An Open-Source Generalist Robot Policy," in *Robotics: Science and Systems (RSS)*, vol. 20, Jul. 2024.

[4] T. L. Team, J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto *et al.*, "A Careful Examination of Large Behavior Models for Multitask Dexterous Manipulation," Jul. 2025, arXiv:2507.05331.

[5] Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu *et al.*, "Qwen2.5 Technical Report," Jan. 2025, arXiv:2412.15115 [cs].

[6] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma *et al.*, "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," Jan. 2025, arXiv:2501.12948 [cs].

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro *et al.*, "LLaMA: Open and Efficient Foundation Language Models," Feb. 2023, arXiv:2302.13971 [cs].

[8] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, Dec. 2023.

[9] OpenAI, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda *et al.*, "GPT-4o System Card," Oct. 2024, arXiv:2410.21276.

[10] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti *et al.*, "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets," Nov. 2023, arXiv:2311.15127.

[11] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang *et al.*, "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer," in *International Conference on Learning Representations (ICLR)*, Oct. 2024.

[12] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu *et al.*, "HunyuanVideo: A Systematic Framework For Large Video Generative Models," Mar. 2025, arXiv:2412.03603 [cs].

[13] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao *et al.*, "Wan: Open and Advanced Large-Scale Video Generative Models," Apr. 2025, arXiv:2503.20314 [cs].

[14] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich *et al.*, "Open-World Object Manipulation using Pre-Trained Vision-Language Models," in *Conference on Robot Learning (CoRL)*, Dec. 2023.

[15] M. Dalal, M. Liu, W. Talbott, C. Chen, D. Pathak, J. Zhang, and R. Salakhutdinov, "Local Policies Enable Zero-shot Long-horizon Manipulation," Mar. 2025, arXiv:2410.22332 [cs].

[16] S. Patel, S. Mohan, H. Mai, U. Jain, S. Lazebnik, and Y. Li, "Robotic Manipulation by Imitating Generated Videos Without Physical Demonstrations," Jul. 2025, arXiv:2507.00990 [cs].

[17] W. Yu, N. Gileadi, C. Fu, S. Kirmani, K.-H. Lee, M. G. Arenas, H.-T. L. Chiang, T. Erez, L. Hasenclever *et al.*, "Language to Rewards for Robotic Skill Synthesis," in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 374–404, iSSN: 2640-3498.

[18] K. Goldberg, "Good old-fashioned engineering can close the 100,000-year "data gap" in robotics," *Science Robotics*, vol. 10, no. 105, p. eaea7390, Aug. 2025.

[19] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-Shot Robotic Manipulation with Pre-Trained Image-Editing Diffusion Models," in *International Conference on Learning Representations (ICLR)*, Oct. 2023.

[20] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani, "Towards Generalizable Zero-Shot Manipulation via Translating Human Interaction Plans," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2024.

[21] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei, "ReKep: Spatio-Temporal Reasoning of Relational Keypoint Constraints for Robotic Manipulation," in *Conference on Robot Learning (CoRL)*, Jan. 2025.

[22] G. Yin, Y. Li, Y. Wang, D. McConachie, P. Shah, K. Hashimoto, H. Zhang, K. Liu, and Y. Li, "CodeDiffuser: Attention-Enhanced Diffusion Policy via VLM-Generated Code for Instruction Ambiguity," in *Robotics: Science and Systems (RSS) XXI*. arXiv, Jun. 2025.

[23] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as Policies: Language Model Programs for Embodied Control," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2023.

[24] M. Xu, Z. Xu, Y. Xu, C. Chi, G. Wetzstein, M. Veloso, and S. Song, "Flow as the Cross-domain Manipulation Interface," in *Conference on Robot Learning (CoRL)*, Jan. 2025.

[25] H. Huang, K. Schmeckpeper, D. Wang, O. Biza, Y. Qian, H. Liu, M. Jia, R. Platt, and R. Walters, "IMAGINATION POLICY: Using Generative Point Cloud Models for Learning Manipulation Policies," in *Conference on Robot Learning (CoRL)*, Sep. 2024.

[26] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning Universal Policies via Text-Guided Video Generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2023.

[27] J. Liang, R. Liu, E. Ozguroglu, S. Sudhakar, A. Dave, P. Tokmakov, S. Song, and C. Vondrick, "Dreamitate: Real-World Visuomotor Policy Learning via Video Generation," in *Conference on Robot Learning (CoRL)*, Jan. 2025.

[28] S. Li, Y. Gao, D. Sadigh, and S. Song, "Unified Video Action Model," Mar. 2025, arXiv:2503.00200 [cs].

[29] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum, "Learning to Act from Actionless Videos through Dense Correspondences," in *International Conference on Learning Representations (ICLR)*, Oct. 2023.

[30] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely, "MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[31] B. Zhang, L. Ke, A. W. Harley, and K. Fragkiadaki, "TAPIP3D: Tracking Any Point in Persistent 3D Geometry," Apr. 2025, arXiv:2504.14717 [cs].

[32] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang *et al.*, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection," in *European Conference on Computer Vision (ECCV)*, 2025.

[33] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland *et al.*, "SAM 2: Segment Anything in Images and Videos," in *International Conference on Learning Representations (ICLR)*, Oct. 2024.

[34] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 32, no. 5, pp. 922–923, Sep. 1976.

[35] H. Jiang, H.-Y. Hsu, K. Zhang, H.-N. Yu, S. Wang, and Y. Li, "PhysTwin: Physics-Informed Reconstruction and Simulation of Deformable Objects from Videos," Mar. 2025, arXiv:2503.17973 [cs].

[36] K. Zhang, B. Li, K. Hauser, and Y. Li, "Particle-Grid Neural Dynamics for Learning Deformable Object Models from RGB-D Videos," in *Robotics: Science and Systems (RSS)*, Jun. 2025.

[37] J. Lee, J. Duan, H. Fang, Y. Deng, S. Liu, B. Li, B. Fang, J. Zhang, Y. R. Wang *et al.*, "MolmoAct: Action Reasoning Models that can Reason in Space," Aug. 2025, arXiv:2508.07917 [cs].

[38] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman *et al.*, "$\pi_0$: A Vision-Language-Action Flow Model for General Robot Control," Nov. 2024, arXiv:2410.24164.

[39] A. Ajay, S. Han, Y. Du, S. Li, A. Gupta, T. Jaakkola, J. Tenenbaum, L. Kaelbling, A. Srivastava, and P. Agrawal, "Compositional Foundation Models for Hierarchical Planning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 22 304–22 325, Dec. 2023.

[40] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani, "Gen2Act: Human Video Generation in Novel Scenarios enables Generalizable Robot Manipulation," Sep. 2024, arXiv:2409.16283.

[41] J. Liang, P. Tokmakov, R. Liu, S. Sudhakar, P. Shah, R. Ambrus, and C. Vondrick, "Video Generators are Robot Policies," Aug. 2025, arXiv:2508.00795 [cs].

[42] C. Yuan, C. Wen, T. Zhang, and Y. Gao, "General Flow as Foundation Affordance for Scalable Robot Learning," Jan. 2024, arXiv:2401.11439 [cs].

[43] B. Eisner, H. Zhang, and D. Held, "FlowBot3D: Learning 3D Articulation Flow to Manipulate Articulated Objects," May 2024, arXiv:2205.04382 [cs].

[44] Z.-H. Yin, S. Yang, and P. Abbeel, "Object-centric 3D Motion Field for Robot Learning from Human Videos," Jun. 2025, arXiv:2506.04227 [cs].

[45] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani, "Track2Act: Predicting Point Tracks from Internet Videos Enables Generalizable Robot Manipulation," in *European Conference on Computer Vision (ECCV)*, 2024.

[46] H. Zhang, B. Eisner, and D. Held, "FlowBot++: Learning Generalized Articulated Objects Manipulation via Articulation Projection," May 2024, arXiv:2306.12893 [cs].

[47] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. Leutenegger, "VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Zero-Shot Robotic Manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[48] H. Zhi, P. Chen, S. Zhou, Y. Dong, Q. Wu, L. Han, and M. Tan, "3DFlowAction: Learning Cross-Embodiment Manipulation from 3D Flow World Model," Jun. 2025, arXiv:2506.06199.

[49] J. Shi, Z. Zhao, T. Wang, I. Pedroza, A. Luo, J. Wang, J. Ma, and D. Jayaraman, "ZeroMimic: Distilling Robotic Manipulation Skills from Web Videos," Mar. 2025, arXiv:2503.23877 [cs].

[50] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning," in *International Conference on Machine Learning (ICML)*, Jun. 2025.

[51] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English *et al.*, "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space," Jun. 2025, arXiv:2506.15742 [cs].

[52] A. Murali, B. Sundaralingam, Y.-W. Chao, W. Yuan, J. Yamada, M. Carlson, F. Ramos, S. Birchfield, D. Fox, and C. Eppner, "GraspGen: A Diffusion-based Framework for 6-DOF Grasping with On-Generator Training," Jul. 2025, arXiv:2507.13097.

[53] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, Oct. 2023.

[54] K. Zhang, B. Li, K. Hauser, and Y. Li, "AdaptiGraph: Material-Adaptive Graph-Based Neural Dynamics for Robotic Manipulation," in *Robotics: Science and Systems (RSS)*, vol. 20, Jul. 2024.

[55] Y. Wang, Y. Li, K. Driggs-Campbell, L. Fei-Fei, and J. Wu, "Dynamic-Resolution Model Learning for Object Pile Manipulation," in *Robotics: Science and Systems (RSS)*, Jul. 2023.

[56] S. Huang, Q. Chen, X. Zhang, J. Sun, and M. Schwager, "ParticleFormer: A 3D Point Cloud World Model for Multi-Object, Multi-Material Robotic Manipulation," in *Conference on Robot Learning (CoRL)*, Jul. 2025.

[57] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. C. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion," vol. 19, Jul. 2023.

[58] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen *et al.*, "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," Jan. 2024, arXiv:2401.14159 [cs].

[59] T. Wiedemer, Y. Li, P. Vicol, S. S. Gu, N. Matarese, K. Swersky, B. Kim, P. Jaini, and R. Geirhos, "Video models are zero-shot learners and reasoners," Sep. 2025, arXiv:2509.20328 [cs].

[60] R. Wang, S. Xu, C. Dai, J. Xiang, Y. Deng, X. Tong, and J. Yang, "MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision," Nov. 2024, arXiv:2410.19115 [cs].

[61] X. Luo, J.-B. Huang, R. Szeliski, K. Matzen, and J. Kopf, "Consistent video depth estimation," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 71:71:1–71:71:13, Aug. 2020.

[62] Z. Zhang, F. Cole, Z. Li, M. Rubinstein, N. Snavely, and W. T. Freeman, "Structure and Motion from Casual Videos," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 20–37.

[63] C. M. Kim, B. Yi, H. Choi, Y. Ma, K. Goldberg, and A. Kanazawa, "PyRoki: A Modular Toolkit for Robot Kinematic Optimization," May 2025, arXiv:2505.03728 [cs].

# Appendices

## A. Video Generation

In this section, we provide additional details on the video generation models, prompt engineering techniques, and the specific prompts used in our experiments.

### A.1. Wan2.1

Wan2.1 is the latest open-source video generation model from Alibaba at the time of this work. We note that a newer version, Wan2.2, was released recently but does not support all the modes of Wan2.1 that we require. Following the official recommendation, we use Chinese prompts, which we found to yield better results than English prompts.

For Wan2.1, we use its Image-to-Video (I2V) model for standard video generation and its First-Last-Frame-to-Video (FLF2V) model when conditioning on a goal image. We generate 41 frames for each video at a resolution of 1280×720 and a frame rate of 16 FPS. We use their UniPC sampler with 40 sampling steps, a noise shift parameter of 5.0, and a guidance scale of 5.0.

### A.2. Veo

We also experimented with Veo, a closed-source model from Google. At the time of our experiments, the model supported I2V generation but not goal-image conditioning. Specifically, we used the `veo-3.0-generate-001` model via the Vertex AI API. We generated 8-second videos[1] at a resolution of 1280×720 and a frame rate of 24 FPS. To maintain consistency with Wan2.1, we downsampled the generated videos to 41 frames.

Pricing for the Veo model is subject to change. At the time of writing, the cost was $0.20 per second of generated video (e.g., $1.60 for an 8-second clip).

### A.3. Prompt Engineering

To improve the quality and controllability of the generated videos, we employ prompt extension, a technique where a simple instruction is automatically enriched with additional details about style, composition, and action.

For Wan2.1, we adapt its official prompt extension script. We use the prompt template from the official repository and pass it to the Gemini 2.5 Pro model to generate the extended Chinese prompt.

For Veo, we utilize its native prompt enhancement feature available through the Vertex AI API, which automatically refines the input prompt for improved generation quality.

### A.4. Generation Prompts

Below, we provide the original and extended prompts used for each task, along with the corresponding initial image from the robot's perspective.

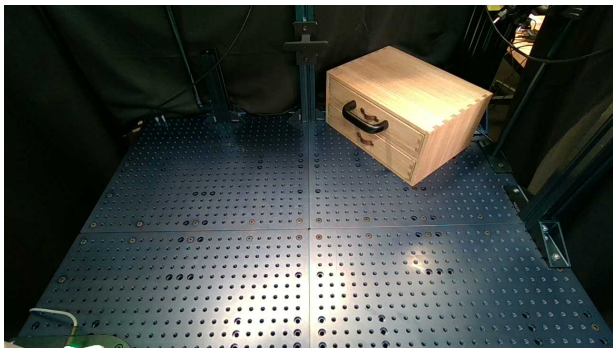

**Figure 8:** Initial observation of the drawer open task.

**Open Drawer (Wan2.1 original)**

一只人手抓住黑色抽屉把手，顺利地将其从抽屉中拉出。抽屉应沿直线打开，且不会前后移动。人手不会在视觉上遮挡抽屉把手。

**Open Drawer (Wan2.1 extended)**

实景拍摄，一只人手抓住黑色抽屉把手，将木质抽屉从木制抽屉箱中顺利拉出。抽屉沿着滑轨直线打开，抽屉把手光滑，人手并未遮挡。整体画面为抽屉箱被固定在带有圆形孔洞的蓝色实验平台上，背景是黑色的幕布和金属支架。展示抽屉打开的全过程。

**Open Drawer (Veo original)**

A human hand grasps a black drawer handle and smoothly pulls the drawer out. The drawer should open in a straight line without moving forward or backward. The human hand should not visually obscure the drawer handle.

[1] During our experiments, the model only supported 8-second video generation. It now also supports durations of 4 and 6 seconds.

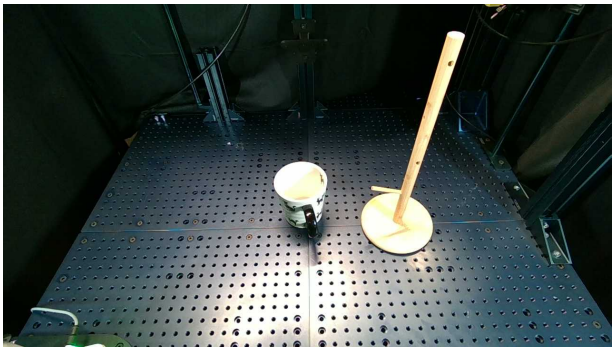

**Figure 9:** Initial observation of the hang mug task.

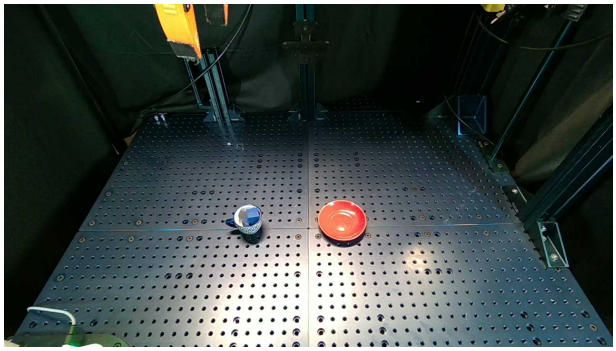

**Figure 10:** Initial observation of the cup on saucer task.

Hang Mug (Wan2.1 original)

一只人的手拿起杯子并将其挂在木架上。人的手不会在视觉上遮挡杯子。

Hang Mug (Wan2.1 extended)

专业工作室摄影，一只手正将一个印有熊猫图案的白色纸杯挂到右侧的木质支架上。纸杯内的液体在过程中有轻微晃动。镜头保持固定，只对物体进行平移，从杯子被拿起，到杯子稳稳地挂在支架上。画面呈现高对比度的冷色调，强调了物品的质感和细节。整体为中景拍摄，突出物体的主体性。

Hang Mug (Veo original)

A human hand picks up the cup and hangs it on the wooden stand. The human hand does not visually obstruct the cup.

Cup on Saucer (Wan2.1 original)

一只人手拿起蓝色的小杯子，举起来，轻轻地放在它红色盘子上。

Cup on Saucer (Wan2.1 extended)

写实风格，一只人手伸出，拿起蓝色带有白色图案的陶瓷小杯子，杯子被稳稳握住。随后，人手将杯子向上举起，并缓慢移动至右侧，将杯子稳稳地放置在一个红色的圆形盘子上。整个过程平稳流畅，镜头从俯视角度拍摄，展现了物体的精细细节和整体的摆放过程。

Cup on Saucer (Veo original)

A human hand picks up the small blue cup, lifts it up, and gently places it on its red plate.

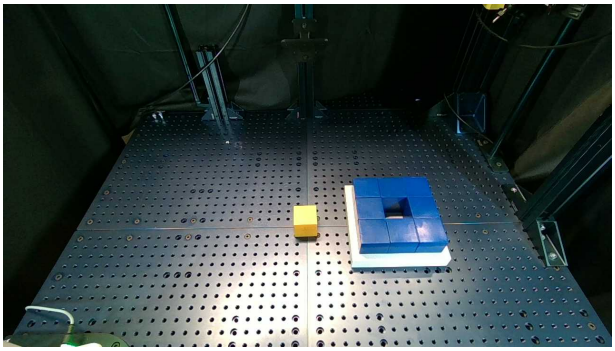

**Figure 11:** Initial observation of the block insertion task.

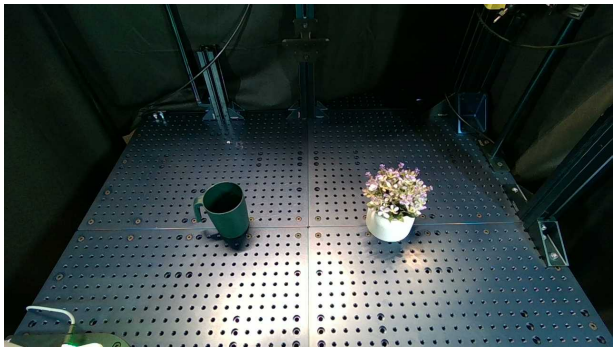

**Figure 12:** Initial observation of the water plant task.

Block Insertion (Wan2.1 original)

一只人手拿起黄色块，并将其准确地插入有蓝色块的盘子的中心。黄色块应该先上升然后下降。人手不会在视觉上遮挡黄色块。

Block Insertion (Wan2.1 extended)

超写实主义近景镜头，一只人手拿起一个黄色方块，然后将它精准地放置在右侧一个由蓝色方块组成的盘子中心。该动作从黄色方块先向上移动，再向下插入盘子中心开始。整个过程中，人手始终保持在黄色方块上方，避免遮挡。画面背景是深色绒布，下方是带有规则孔洞的金属实验平台。镜头稳定，画面清晰。

Block Insertion (Veo original)

A human hand picks up the yellow block and inserts it precisely into the center of the plate with the blue block. The yellow block should rise first and then fall. The human hand should not visually obscure the yellow block.

Water Plant (Wan2.1 original)

一只人手抓住左边的绿色水杯，将其举起，然后平稳地给植物浇水。摄像机始终保持静止。人手不会在视觉上遮挡杯子。

Water Plant (Wan2.1 extended)

仰视视角，一只手持绿色马克杯，马克杯倾斜，准备向白盆里的紫色小花浇水。背景为深邃的黑色布景，地面为带有网格的金属平台，光线集中在平台中央，形成明暗对比。整体画面呈现出一种科技感和实验性。

Water Plant (Veo original)

A human hand grasps the green water cup on the left, lifts it, and steadily waters the plant. The camera remains stationary throughout. The hand does not visually obscure the cup.

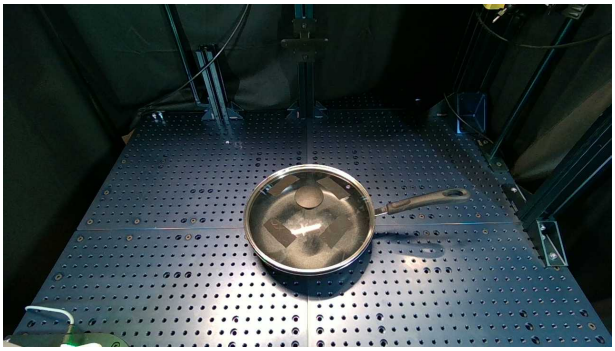

**Figure 13:** Initial observation of the open lid task.

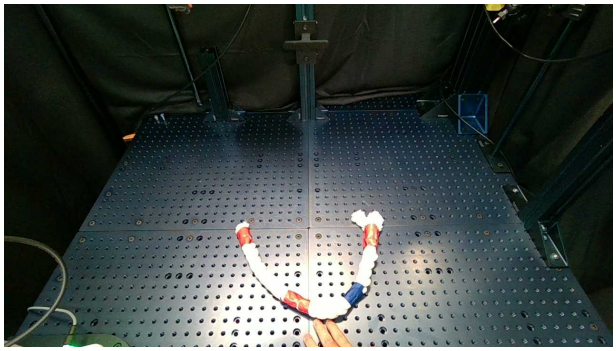

**Figure 14:** Initial observation of the straighten rope task.

Open Lid (Wan2.1 original)

一只人手抓住锅的透明盖子并将其直接提起。摄像机始终保持静止。人手不会在视觉上遮挡盖子。

Open Lid (Wan2.1 extended)

写实主义摄影，一只手在抓住平底锅的透明锅盖，并将其提起。锅盖上贴有三张长方形的黑色不干胶，上面有白色的反光。锅盖中间有一个黑色的圆形把手。锅的四周是深蓝色的实验台，表面布满了规则排列的圆孔。背景是深色的幕布和金属支架，光线集中在锅具上，营造出一种局部照明的氛围。近景，固定镜头，俯视角度。

Open Lid (Veo original)

A human hand grasps the transparent lid of a pot and lifts it straight up. The camera remains stationary. The hand does not visually obscure the lid.

Straighten Rope (Wan2.1 original)

一只人手缓缓地把弯曲的绳子推成直的。

Straighten Rope (Wan2.1 extended)

工业风写实记录，在布满孔洞的深色金属实验台上，一只戴着黑色智能手表的人手缓缓出现，将一根红白蓝三色相间的粗布绳从U形弯曲状态，慢慢地推动、抚平成一条直线。整个动作流畅而稳定。镜头固定，采用俯视视角，记录了这一精准的操作过程。

Straighten Rope (Veo original)

A hand slowly pushes the bent rope into a straight line.

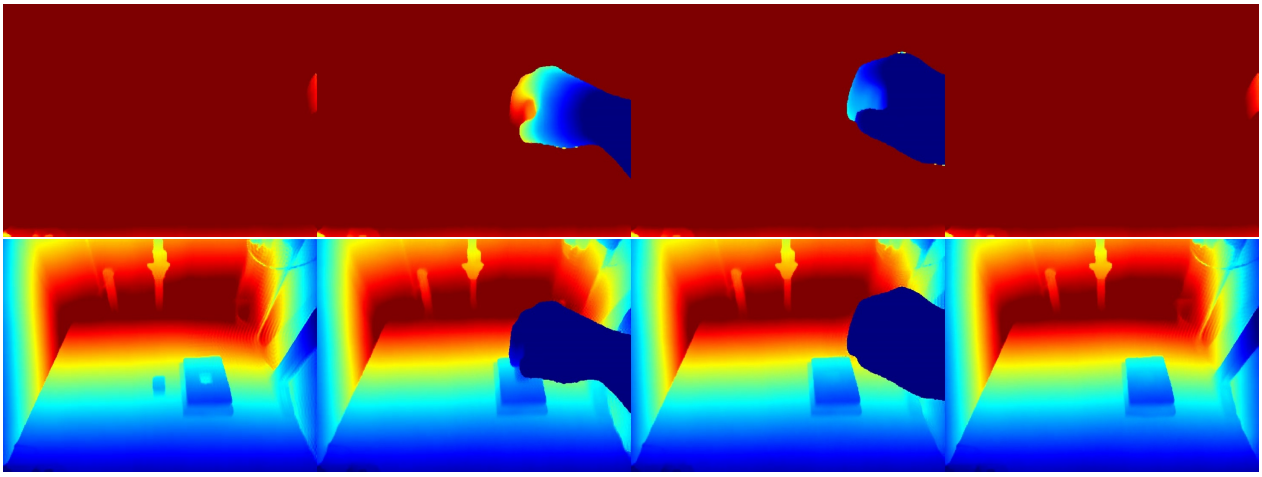

**Figure 15: Comparison of the depth maps before (top) and after (bottom) scaling**. The depth maps are visualized in the `jet` color map, where the colormap range is obtained from the ground-truth depth map.

## B. Depth Estimation

We use the implementation of MegaSaM [30] from TAPIP3D [31] for depth estimation. Specifically, we use the MoGe [60] model to estimate the per-frame depth map. Instead of estimating the camera intrinsics, we use the ground-truth calibrated intrinsics from our camera. The estimated depth maps are then postprocessed by bundle adjustment and consistent video depth (CVD) [61] optimization following CausalSAM [62].

Even after these postprocessing steps, the estimated metric depth maps are still ambiguous. Therefore, we opt to use the initial ground-truth depth map as the reference depth map for calibration. Specifically, we compute the scaling factor between the median depth of the first estimated frame and the initial ground-truth depth map. We then multiply the estimated depth maps by this scaling factor to obtain the calibrated depth maps.

After applying the scaling factor (see Fig. 15), the calibrated depth maps are more consistent with the ground-truth depth map and more temporally consistent. We find the calibrated depth maps are accurate enough to support precise manipulation tasks such as the block insertion as demonstrated in the figure, which requires millimeter-level precision.

## C. 3D Point Tracking

We leverage TAPIP3D [31] for 3D point tracking, which tracks 3D points in the XYZ 3D coordinate space instead of the UVD 2D space. We generate query points on the first frame using uniform grid sampling of $32 \times 32$ points. We set the tracker iterations to 6.

## D. Object Grounding

The previous step produces dense 3D tracking for the entire image. We need to ground the points to the target object. We use the Grounded-SAM2 pipeline [58] for object grounding, which combines Grounding DINO [32] and SAM2 [33]. We pass the query object name to the pipeline to extract the object mask throughout the video. Then, we use the mask video to filter the 3D tracking points and only keep the points that are visible throughout the video. We set the bounding box threshold to 0.25 and text threshold to 0.3. We select the bounding box from Grounding DINO with the highest score and set it as the input prompt to SAM2 to extract the object mask.

## E. Rejection Sampling

After obtaining the 3D object flow, we can project the 3D flow onto the first frame to obtain the 2D object flow. We then pass object flow images to Gemini 2.5 Pro to filter out hallucinations, such as generative artifacts and implausible motions, that may be unavoidably introduced by the video generation model. We find this strategy to be effective and benefits from scaling the execution-time computation resources. For example, we can generate 8 candidates in parallel and select the best one from the 8 candidates using 8 H100 GPUs.

Along with the following system prompt, we also provide the flow image and the task description used to generate the video to clarify our expectation of

the task.

**Rejection Sampling System Prompt**

You are a flow analysis expert. Analyze the stitched flow image and evaluate which flow visualization (marked with IDs in the top-left corner) represents the most reasonable and natural object motion. The flow is a manual annotation overlay on the image to indicate the intended object motion. Consider: 1. Continuity and smoothness of the flow 2. Natural motion patterns 3. Proper object identification (avoid flows that spread throughout the entire image) 4. Alignment with the task requirements You should reject images that show flows throughout the image (which means the object is not identified). Provide a clear recommendation on which flow ID is best and why.

## F. Trajectory Optimization

During execution time, we refine the sequence of actions using trajectory optimization to find an optimal, collision-free, and smooth sequence of joint configurations $Q = \{q_0, q_1, \dots, q_{T-1}\}$. The trajectory is initialized by linearly interpolating between start and end configurations, $q_{\text{start,IK}}$ and $q_{\text{end,IK}}$, which are pre-calculated using an IK solver. The optimal trajectory $Q^*$ is found by solving the following constrained non-linear optimization problem:

$$\begin{aligned} \min_{Q} \quad & w_s C_{\text{smooth}} + w_r C_{\text{rest}}, \quad \text{subject to} \\ & q_0 = q_{\text{start,IK}} \quad \text{and} \quad q_{T-1} = q_{\text{end,IK}}, \\ & q_{\min} \le q_t \le q_{\max}, \quad \forall t \in \{0, \dots, T-1\}, \\ & d_s(q_t, q_{t+1}, O_j) \ge \epsilon_{\text{safe}}, \quad \forall t, \forall O_j \in \text{Obstacles}. \end{aligned} \tag{6}$$

This optimization problem is solved using a non-linear least-squares algorithm (Levenberg-Marquardt). The constraints for joint limits and collision avoidance are incorporated as high-weight penalty terms in the objective function, while start and end configurations are hard constraints.

### Objective and Penalty Terms

**Smoothness Cost ($C_{\text{smooth}}$).** This cost penalizes non-smooth motion by minimizing the squared norms of joint velocity ($\dot{q}$), which is approximated using finite differences. In practice, this is implemented as a cost on the deviation from the previous joint configuration, encouraging temporal smoothness:

$$C_{\text{smooth}} = \sum_t w_s \|q_t - q_{t-1}\|^2. \tag{7}$$

**Rest Position Cost ($C_{\text{rest}}$).** This is a regularization term that encourages the trajectory to remain close to a default home configuration, $q_{\text{rest}}$:

$$C_{\text{rest}} = \sum_t w_r \|q_t - q_{\text{rest}}\|^2. \tag{8}$$

**Joint Limits Penalty**. This term penalizes any violation of the minimum ($q_{\min}$) and maximum ($q_{\max}$) joint limits:

$$\begin{aligned} C_{\text{limits}} = \sum_t w_l \big( & \| \max(0, q_t - q_{\max}) \|^2 \\ & + \| \max(0, q_{\min} - q_t) \|^2 \big). \end{aligned} \tag{9}$$

**Collision Avoidance Penalty**. This term enforces a safety margin, $\epsilon_{\text{safe}}$, from world obstacles, $O_j$, by applying a hinge loss on the signed distance, $d_s$, of the robot's swept volume:

$$C_{\text{collision}} = \sum_{t,j} w_c \cdot \max(0, \epsilon_{\text{safe}} - d_s(q_t, q_{t+1}, O_j))^2. \tag{10}$$

We implement the optimization using PyRoki [63] and Jax. The cost terms are weighted as follows. The joint limit penalty is set to a high value of $w_l = 100.0$ to act as a hard constraint. The smoothness weight is $w_s = 10.0$, the collision penalty weight is $w_c = 15.0$, and a small regularization is applied with a rest pose weight of $w_r = 0.1$.

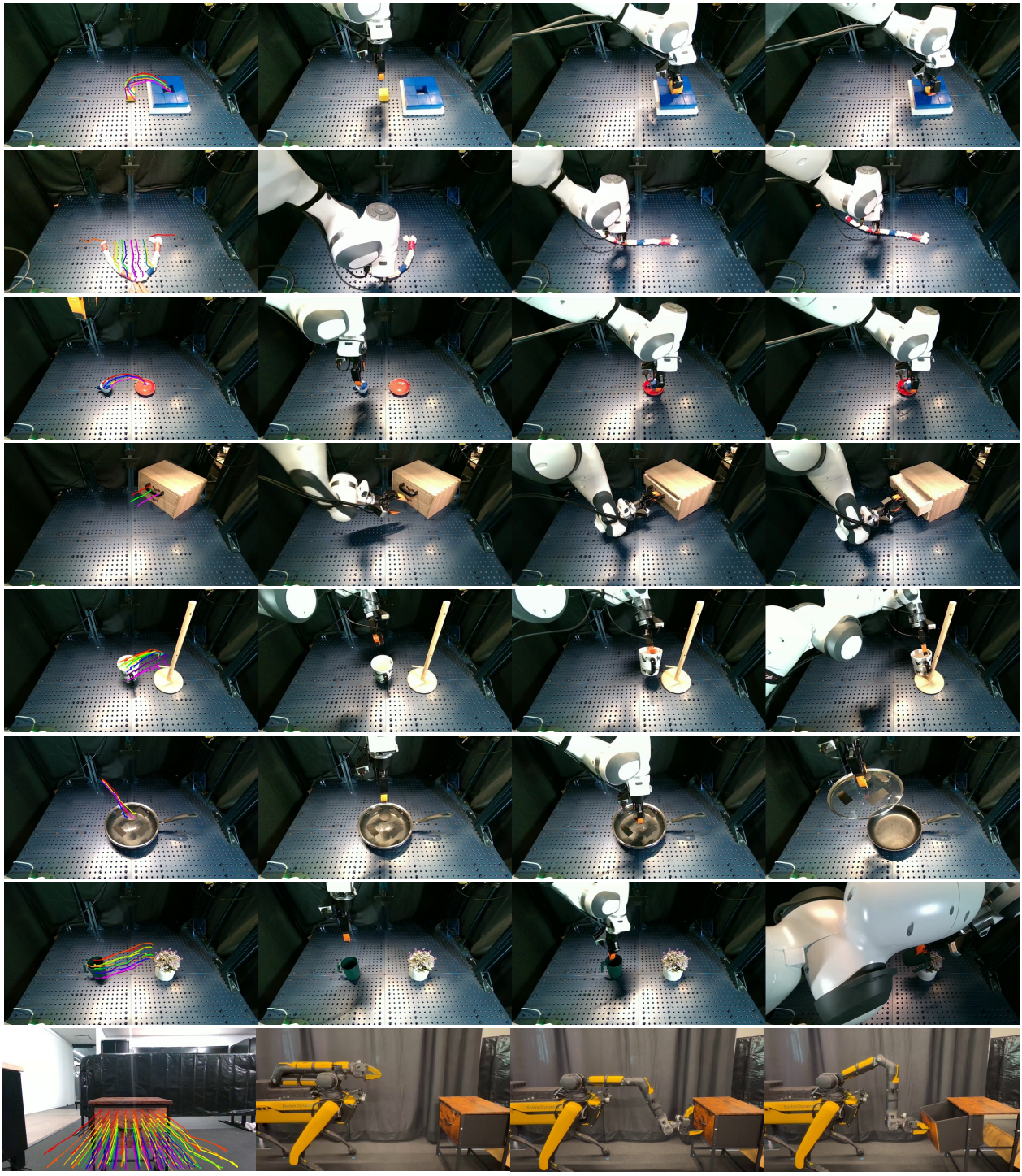

**Figure 16: Real-world manipulation experiments.** From top to bottom: block insertion, rope straightening, cup on saucer, open drawer, hang mug, open lid, water plant, and open drawer using the Spot.

## G. Experiments

We show more visualizations of the real-world manipulation experiments in Fig. 16. NovaFlow is versatile and supports cross-embodiment manipulation, which we use to manipulate rigid, deformable, and articulated objects using tabletop and mobile manipulator. It is also viewpoint-agnostic and can be deployed on a novel platform after performing hand-eye calibration.