CoDex: Learning Compositional Dexterous Functional Manipulation without Demonstrations

Anonymous

***Abstract*— Functional Object Manipulation (FOM) tasks require interacting with an object to elicit its intended function. These tasks span from simple object usage, such as hammering a nail, to more complex interactions like spraying a plant with a water sprayer. Often, these complex interactions feature internal object mechanisms such as triggers or buttons, and the task success requires directing the effect of the actuation to other objects in the scene, e.g., when aiming and actuating a spray bottle or a glue gun. *Compositional Dexterous FOM* tasks require coordinated control over both the object's internal and external degrees of freedom, posing a significant challenge for robots due to the demanding integration of semantic understanding (of the object's function, actuation mode, and application area) with intricate physical dexterity (to manage grasp stability, movement trajectory, and actuation). We introduce CoDex, a zero-demonstration framework leveraging vision-language models (VLM) to generate *semantic constraints* and enforcing them via *analytic constrained optimization*, and *constraint-guided RL* to compose grasp–move–actuate behaviors that transfer directly from simulation to the real world. We evaluate CoDex to control a 7 DoF robot arm with a 16 DoF multifingered hand in six Compositional Dexterous FOM tasks involving previously unseen objects with internal mechanisms(spray bottles, hot glue gun, air duster) and their application on various unseen target objects, showcasing its ability to autonomously discover and execute complex, physically viable dexterous behaviors without human demonstrations. More information at `https://codex-2025.github.io/`.**

I. Introduction

Imagine a robot tasked with spraying a plant: it needs to grasp the bottle stably, move and aim it correctly towards the leaves, and squeeze the trigger to activate its functionality, all in a coordinated sequence. This kind of task is a special case of Functional Object Manipulation (FOM) [1]–[5], where for the agent to use the object for its specific purpose. It must (1) actuate the object's internal degrees of freedom (e.g., trigger, button, lever) while (2) coordinating control of its external DoF (object pose) to apply the function to the intended target region [6], [7]. Due to their high demands on internal-external DoF coordination and mechanism actuation, these **Compositional Dexterous Functional Object Manipulation (CD-FOM)** tasks remain an open challenge in robotics.

CD-FOM requires bridging the gap between semantic understanding and intricate physical dexterity. The robot must not only interpret the task context—understand the object's function, identify how and where to interact with it to actuate it (*local* semantics), and reason about the desired outcome relative to other objects (*global* semantics)—but also be able to execute the task physically—achieve a stable yet functional grasp, coordinate complex hand-arm motions, and apply precise forces. Effectively leveraging semantic reasoning to guide the physical skills is crucial for success.

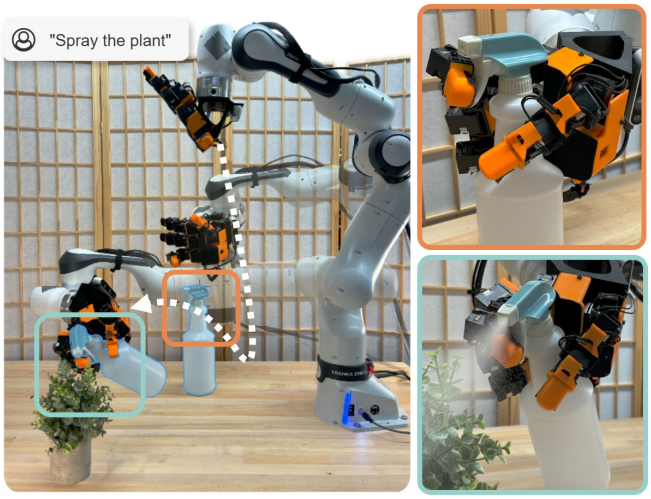

Fig. 1: Our method, **CoDex**, bridges high-level semantic understanding with low-level physical dexterity. The visualization shows CoDex executing the task "*spray the plant*" from a language command. The robot autonomously performs a task-aware functional grasp, repositions the spray bottle, and actuates the trigger to spray mist, all without requiring human demonstrations.

General object manipulation methods fall short for CD-FOM. Learning from Demonstration methods acquire dexterity from expert teachers, while semantic understanding is implicitly captured in their behavior [8]–[11]. However, learning the correlations between semantics and dexterity from demonstrations is hard to obtain since it requires large amounts of data obtained by teleoperating complex, multi-fingered hands to actuate objects with internal mechanisms [10]–[13]. Recent imitation from human videos methods eliminate the need for labor-intensive teleoperation, but require instead learning to overcome human-robot morphological differences with limited object-specific strategies [7], [14]–[16]. Alternatively, optimization-based approaches such as Reinforcement Learning [17], [18] and analytical grasp synthesis [19]–[23] achieve CD-FOM physical dexterity without demonstrations, but their lack of semantic understanding demands external object-specific guidance in the form of reward design and optimization objectives, which limits their applicability and autonomy [2], [5], [24]–[26]. Such general semantic understanding can be obtained from large-scale pre-trained models, such as Vision-Language Models (VLMs) [27]–[31], but their initial integration into robotic solutions [30], [31] revealed their limitations in geometric and embodied understanding, which restricted their

guidance so far to a coarse, abstract level, not sufficient for the intricate, coordinated hand-and-arm motions required for CD-FOM [27]–[29].

In this work we introduce **CoDex**, a framework that bridges semantic understanding and physical dexterity for CD-FOM through the use of *semantic constraints*. We define semantic constraints as a set of geometric and spatial conditions derived from an object's function and the overall task goal. CoDex integrates a VLM into an iterative refinement procedure to achieve zero-demonstration semantic understanding, interpreting the task and generating two types of semantic constraints: local (e.g., where to press a trigger and in which direction) and global (e.g., where to aim a nozzle). These constraints are used by CoDex to achieve physical dexterity, guiding a two-step policy learning process: an analytic constrained optimization step that synthesizes a diverse set of functionally-aligned grasp candidates, and a constraint-guided RL step that uses the synthesized grasps and the semantic constraints as guidance to train a sensorimotor arm-and-hand policy for the CD-FOM task.

We demonstrate the capabilities of CoDex to operate six previously unseen objects for different CD-FOM tasks, controlling a 7-DoF robot arm equipped with a 16-DoF multi-fingered hand and achieving 73% average combined success rate. Our experiments validate that the VLM-generated semantic constraints are crucial to this performance. A human participant study demonstrates that our method for determining global constraints produces significantly more appropriate poses than prior VLM-based approaches. Furthermore, we show that our final policy learning stage is critical for achieving physical dexterity, improving functional success by over 40% when compared to using analytical grasps combined with the same VLM-generated constraints.

II. Related Work

CoDex bridges semantic understanding and physical dexterity for compositional dexterous functional object manipulation (CD-FOM). We position our work relative to three key research areas.

Semantic Understanding via Vision-Language Models. Vision-Language Models (VLMs) provide zero-shot semantic understanding for robotics [27]–[31], interpreting high-level goals from language and vision. However, VLMs typically provide abstract guidance lacking the detailed physical understanding needed for dexterous tasks [27]–[29]. Recent VLM-based manipulation systems [28], [29], [32] show promise but either require extensive training data or remain limited to coarse manipulations. ReKep [31] uses VLMs to generate keypoint constraints for manipulation, while PIVOT [30] employs iterative visual prompting to refine robot actions. CoDex leverages VLMs to generate concrete semantic constraints—both local (actuation and function points) and global (target poses)—that directly inform constrained optimization and policy learning, effectively translating abstract VLM knowledge into CD-FOM policies without requiring task-specific training data.

Functional Object Grasping and Physical Dexterity. In task-oriented grasping, the aim is to select a grasp that not only stabilizes the object but also facilitates the intended function [4], [5], [25], [26], [33]–[35]. Several methods focus on optimizing contacts to achieve stable grasps, often predicting force-closure metrics like the Ferrari-Canny [25], [34]–[36], but cannot be applied to grasps that enable the actuation of objects' internal degrees of freedom. Recent analytical strategies [4], [26], [33] consider internal degrees of freedom but they do not compose them with post-grasping trajectories to actuate the object at the right location to achieve a task. All these methods aim to provide a grasp synthesis solution that generates a successful grasp from images to be executed by a predefined controller, which can lead to failures. Recently, some methods have integrated a simulator with a model of the specific object into the loop for online improvement of grasping strategies using reinforcement learning or exploiting the simulator's differentiability for optimization [2], [25], [37]–[39]. While demonstrating better performance, this strategy requires manual annotation of the rewards and has yet to be extended to complex objects with internal degrees of freedom (DoF) and post-grasping motion. Moreover, all previous methods improve grasp stability and/or functionality, but treat grasping as an isolated problem, missing the opportunity to reason about it in conjunction with subsequent motion to enhance dynamics and stability during actuation.

Compositional Dexterous Functional Object Manipulation. Often, successful tool use demands composing the control of both in-hand adjustments and whole-arm extrinsic motions [6], [7], [24], [40], but most existing works on dexterous manipulation focus on one or the other. For instance, significant research addresses in-hand manipulation, focusing on fine finger coordination for tasks like object reorientation [5], [13], [24], [39], or rotating caps [2], [41], [42], without consideration for full arm motion to control the object's overall trajectory for a task. Other works tackle extrinsic manipulation, where a grasped tool interacts with the environment, such as hammering or shoveling [5], [43], [44] on objects without internal degrees of freedom. Closer to ours, [5] provide a composed solution for arm and multi-fingered hand motion, but their method requires human demonstrations and focuses on optimizing the grasp stability to resist the forces resulting from subsequent arm motion, failing to actuate the object's internal degrees of freedom. In contrast, CoDex holistically addresses the entire grasp-move-and-actuate problem, generating a composed solution that actuates both the object's internal and external degrees of freedom.

III. CoDex: Compositional Dexterous Functional Object Manipulation

CoDex bridges semantic understanding and physical dexterity by leveraging VLM outputs as semantic constraints, which are then enforced through analytic constrained optimization and constraint-guided RL. As illustrated in Fig. 2, given a language task description $\mathcal{L}$ and

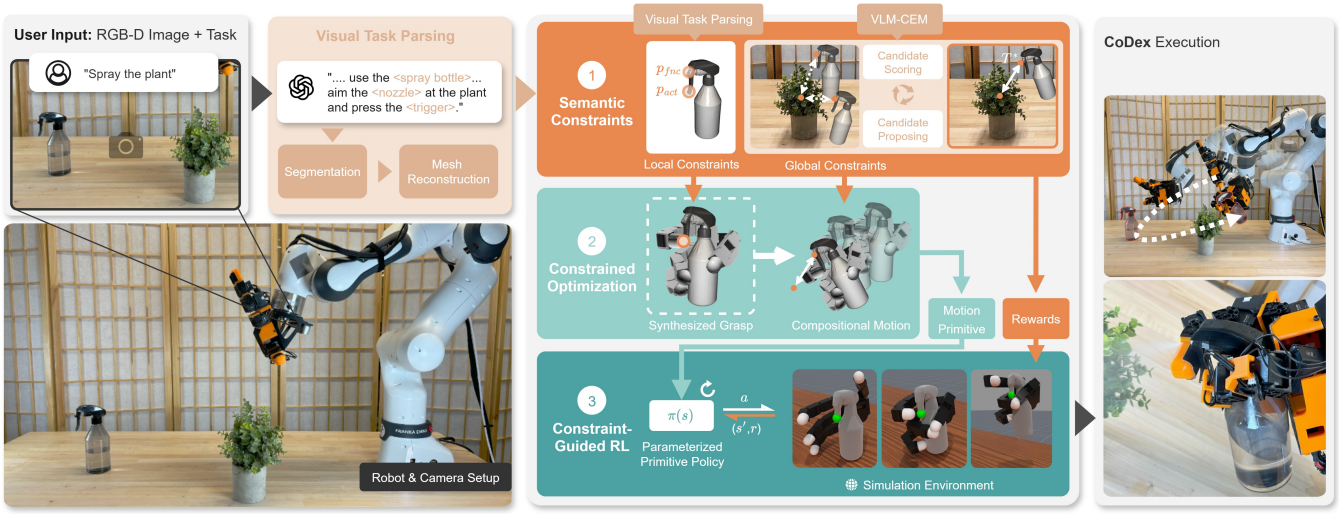

Fig. 2: Overview of the CoDex pipeline, which bridges high-level VLM understanding and low-level dexterity by translating abstract VLM outputs into concrete semantic constraints that guide a two-stage policy learning process. (1) **VLM-Generated Semantic Constraints.** First, a VLM interprets the user's input to generate local constraints (key interaction points like the actuation point and function point) and a global constraint (the final object pose). (2) **Constrained Optimization.** In the first learning phase, these constraints are enforced through analytic constrained optimization to synthesize a diverse set of motion trajectories, each of which includes a task-aware functional grasp. (3) **RL Policy Training.** In the second phase, these motion trajectories initialize a constraint-guided RL process, which uses the same semantic constraints as reward function and learn the complete grasp-move-actuate policy.

Fig. 3: Reconstructed objects with their VLM-identified **local semantic constraints**. To extract these points, a VLM first generates text descriptions for the actuation and function points. MoLMo [45] grounds these descriptions to 2D pixel coordinates (u, v) on the input image. To project these into 3D, we first align the reconstructed mesh to the image using FoundationPose. The 2D points are then unprojected onto the 3D mesh using depth information to yield the final actuation point p_{act} (blue arrow start) and function point p_{fnc} (orange arrow start). The actuation direction d_{act} and function direction d_{fnc} are estimated as the negative surface normal (dotted line indicates *inside the object*).

an RGB-D scene observation I, our pipeline sequentially executes two stages: VLM-Generated Semantic Constraints and Constraint-Guided Policy Training. Constraint-guided Policy Training can be further divided into two sub-stages: Constrained Optimization and Constraint-Guided RL. Below, we detail each stage.

A. VLM-Generated Semantic Constraints

Given the input pair $(\mathcal{L}, I)$, this stage uses VLMs to generate a set of semantic constraints that guide policy learning. These constraints include: (i) **local semantic constraints**, the actuation point p_{act} and function point p_{fnc} on the object, and the target point p_{tgt} on the environment; and (ii) a **global semantic constraint**, the object's target 6D pose T^*.

The process starts with *Visual Task Parsing*. The functional object is identified from $(\mathcal{L}, I)$ using open-vocabulary segmentation (LangSAM). The segmented functional object's 3D mesh $\mathcal{M}$ is then constructed using Tripo [46].

1) Local Semantic Constraints: Next, to derive the local semantic constraints, VLM queries generate text descriptions of key interaction points (e.g., "trigger", "nozzle"). These descriptions are then grounded to 2D pixel coordinates (u, v) on the image, which are subsequently unprojected into 3D space to identify the **Actuation Point** p_{act} and **Function Point** p_{fnc} on the mesh, as detailed in Fig. 3. The actuation direction d_{act} is approximated by the negative surface normal at p_{act}.

2) Global Semantic Constraints: Finally, CoDex derives **global semantic constraints** (the goal pose T^*) through *VLM-Guided Cross-Entropy Method* (VLM-CEM), an algorithm inspired by [30]. VLM-CEM leverages the VLM's reasoning to drive an iterative pose search: at each round, the VLM is prompted with the history $\mathcal{H}$ of previously scored candidates and proposes K new poses expected to score higher. We render these candidates (see Fig. 9), score them with the VLM, append them to $\mathcal{H}$, keep only an elite subset, and repeat for N rounds. As detailed in Algorithm 1, this iterative proposal-and-evaluation loop allows the VLM to perform an implicit optimization, converging on a pose that is both functionally correct and physically grounded.

B. Constraint-Guided Policy Training

Using the semantic constraints generated in the previous stage, this stage learns a policy π for the complete grasp-move-actuate sequence. This is achieved through a two-phase process: (1) Analytic Constrained Optimization generates a diverse set of statically stable and functionally viable motions, and (2) Constraint-Guided RL uses them as initialization for an online reinforcement learning process with action primitives in simulation.

Algorithm 1 VLM-CEM

Require: Function Point p_{fnc}, Target Location p_{tgt}, iterations N=6, candidates K=10
Ensure: goal pose T^*

$T_0 \leftarrow \textsc{AnchorInit}(p_{\text{fnc}}, p_{\text{tgt}})$
$s_0 \leftarrow \textsc{VLMScore}(\textsc{Render}(T_0))$
$\mathcal{H} \leftarrow \{(T_0, s_0)\}$

for $i = 1$ **to** N **do**
  $\mathcal{P} \leftarrow \textsc{VLMPropose}(\mathcal{H}, K)$
  for each $T \in \mathcal{P}$ **do**
    $s \leftarrow \textsc{VLMScore}(\textsc{Render}(T))$
    $\mathcal{H} \leftarrow \mathcal{H} \cup \{(T, s)\}$
  end for
  $\mathcal{H} \leftarrow \textsc{TopR}(\mathcal{H}, \rho)$ ▷ keep top-ρ elites (optional)
end for
return $T^* \leftarrow \arg\max_{(T,s)\in\mathcal{H}} s$

Fig. 4: Human-like (*left*) and robot-specific (*right*) examples of initial functional grasp candidates. Our analytic constrained optimization synthesizes functionally valid human-like and robot-specific grasps allowing CoDex to exploit the hand's full morphology instead of restricting it to the human grasps that can be obtained with imitation learning.

1) Analytic Constrained Optimization: This phase translates the VLM-generated local semantic constraints into concrete mathematical objectives for grasp synthesis. We first sample initial hand configurations q_0 from $\mathcal{Q}$—the valid joint space—using inverse kinematics, biasing a finger to be near the actuation point p_{act}. These samples are then refined via constrained optimization. The local constraints ($p_{\text{act}}, d_{\text{act}}$) directly inform the functional terms in Eq. 1: the *Actuation Pt Proximity* term ensures a fingertip is within a distance δ_{dist} of p_{act}, while the *Actuation Alignment* term ensures the fingerpad normal aligns with $-d_{\text{act}}$. These are optimized alongside physical constraints (stability, collision avoidance) and an objective to maximize grasp robustness, measured by the min-weight force closure metric $l^*(q)$ [23], [36].

$$
\begin{aligned}
\max_{q\in\mathcal{Q}} \quad & l^*(q) \\
\text{s.t.} \quad & l^*(q) \geq l_{\min} && \text{(Min F. Closure)} \\
& s(FK_i(q)) = 0,\ \forall i \in \{1, ..., n_c\} && \text{(Surface Contact)} \\
& \sigma_j(q) \geq d_j,\ \forall \text{ collision pairs } j && \text{(Coll. Avoidance)} \\
& \exists i_{act} \in F_{act} \text{ s.t.} \\
& \quad \|FK_{tip}(q, i_{act}) - p_{act}\|_2 \leq \delta_{dist} && \text{(Act Pt Proximity)} \\
& \quad n_{pad}(q, i_{act}) \cdot (-d_{act}) \geq \cos(\delta_{angle}) && \text{(Act Alignment)}
\end{aligned}
\tag{1}
$$

Fig. 5: Six functional object manipulation tasks in our experiments. They require combining local manipulation of functional objects with internal DoF (flashlight, board spray, water spray, air blower, hot glue gun, and salt grinder) with their global motion in the scene.

If the optimization fails, we resample q_0 and restart. This process yields a diverse set of feasible, function-aligned candidates (Fig. 4) for initializing the RL policy.

2) Constraint-Guided RL: While the analytic candidates provide functionally-aligned and statically stable starting points, they do not account for the dynamics of the full manipulation task. The goal of this stage is therefore to learn a policy π that discovers a complete, dynamically robust motion sequence for the entire grasp-move-actuate task. By initializing the RL policy with the optimized candidates, we significantly constrain the exploration problem, enabling the agent to focus on learning the complex dynamics of contact, movement, and actuation.

Policy Parametrization The learned policy π takes an observation o consisting of the relative hand pose, candidate finger joints, and the designated actuation finger index. It outputs an action vector $a \in \mathbb{R}^{38}$ that parameterizes a motion primitive by defining targets relative to the input candidate q_{cand}:

- Target Grasp Joint Offsets: $\Delta j_{\text{grasp}} \in \mathbb{R}^{16}$.
- Target Pre-Grasp Joint Percentages: $p_{\text{pregrasp}} \in \mathbb{R}^{16}$.
- Residual Hand Pose Offset: $\Delta T_{\text{hand}} \in \mathbb{R}^{6}$.

Motion Primitive. These action parameters guide a multi-stage motion primitive, visualized in Fig. 6. The primitive breaks down the complex task into a sequence of simpler steps: (1) the hand first moves to a pre-contact pose near the object, (2) it then transitions to the final grasp pose, (3) the fingers close to secure the object, (4) finally, it moves the object towards the global goal T^* while simultaneously performing the required actuation. This structured, parameterized primitive makes the high-dimensional control problem tractable for RL.

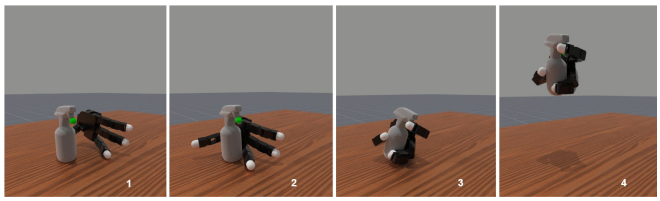

Fig. 6: Key stages of the CoDex's parameterized motion primitive trained in simulation. The policy action space determines (1) the pre-contact approach, (2) grasp pose, (3) finger closing strategy (internal DoF actuation), and (4) object pose change (external DoF actuation).

Reward Function The policy is trained with PPO to maximize a unified reward function R. The reward is a normalized weighted sum of shaped rewards (R_k) and binary stage success flags (S_k), plus a bonus for final success (S_{success}). A penalty resets the reward to 0 if significant grasp instability is detected. The shaped rewards (R_k) directly enforce the semantic constraints by providing dense feedback for maintaining alignment with ($p_{\text{act}}, d_{\text{act}}$) and applying sufficient force along the actuation direction. This unified structure avoids the need for task-specific reward engineering [26]. The entire online training process converges in approximately one hour in MANISKILL3 simulation [47].

The result of the RL is a full policy that generates a compositional grasp-move-actuate motion for the CD-FOM task. In the final step, the RL policy is executed on the real robot, using a Franka arm with a LEAP hand.

IV. EXPERIMENTAL EVALUATION

In our experiments, we evaluate whether the proposed **CoDex** framework successfully bridges the gap between high–level vision–language understanding and low-level, physics-grounded execution. To that end, our experiments aim to answer the following three research questions:

Q1 *How well does **CoDex** perform in CD-FOM in the real world?*

Q2 *Does the **VLM-CEM** procedure propose global constraints that are both semantically **and** physically appropriate for the task?*

Q3 *How much does the **constraint-guided RL** improve success compared to attempting directly the candidate grasps from constrained optimization?*

Experimental Setup: We test CoDex on a 7-DoF FRANKA Emika Panda arm, with a 16-DoF LEAP Hand mounted as end-effector, as shown in Fig 5. RL Policies are trained in MANISKILL3 using 2,048 parallel worlds and directly deployed on the real robot. We test on the six functional manipulation tasks introduced in §III. At the start of every episode, the object is placed at a random pose on the table before starting the trial. Each trial is evaluated with two binary criteria: (i) Correct object movement–the object is moved to the task-required pose, (ii) Correct actuation–the mechanism is triggered. A trial is considered success if both the criteria are fulfilled and there are no other failures. Additionally, we record the grasp stability (i.e., the object does not slip from the hand) for analysis in Q3.

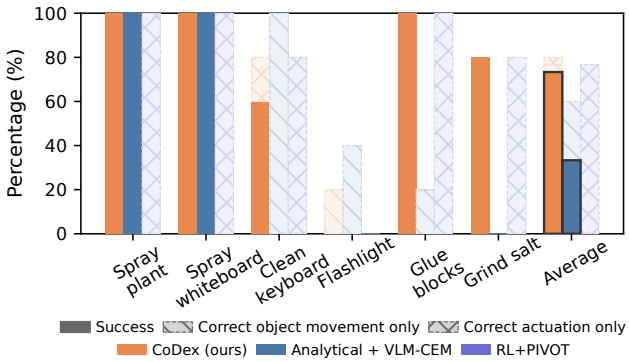

Fig. 7: Overall success rate comparison on six CD-FOM tasks, evaluated over five real-world trials per task. Solid segments represent the rate of **success**, while shaded segments show rates of **partial success** (either correct object movement only or correct internal DoF actuation only). CoDex achieves a 73% success rate, demonstrating the significant benefit of its policy learning stage and VLM-CEM compared to two baselines, pure analytical grasp synthesis [23] and PiVOT [30].

Q1 — Performance on Compositional Dexterous Functional Manipulation Tasks

Figure 7 summarizes our real-world results, comparing CoDex to two baselines: *Analytical+VLM-CEM* and *RL+PIVOT*. Our method achieves an overall task success rate of **73%**, significantly outperforming the baseline's **33%** and **0%**. The *Analytical+VLM-CEM* baseline combines the best-performing initial grasp candidate from Li et al. [23] (selected via an oracle for maximum stability, see Q3) with the global constraint from our VLM-CEM. The *RL+PIVOT* baseline uses the same policy training procedure as ours but uses PIVOT [30] (SE3 variant) to generate global constraints. It has a 0% success rate because none of PIVOT's generated global constraint meets the task requirement (e.g., the spray bottle is not correctly aimed at the plant). We further explore this in Q2.

The performance gap between our method and *Analytical+VLM-CEM* primarily stems from actuation failures in the baseline. While the oracle-selected grasp is stable for lifting, the object often shifts slightly within the hand during the subsequent movement to the goal. This minor slippage is frequently enough to misalign the actuation finger, causing it to lose the precise contact required to operate the object's mechanism. This highlights a core challenge of CD-FOM: a statically optimal grasp is often insufficient for a dynamic, contact-rich sequence. The policy learning procedure in CoDex is crucial because it learns to actively maintain and adjust the grasp throughout the entire composed motion for the final functional usage of the object.

Failures of CoDex, concentrated primarily on the `illuminate toy`, `clean keyboard`, and `grind salt` tasks, highlight complexities in integrating semantic goals with physical execution for these demanding FOM tasks.

A common failure mode arises from *grasp–placement*

incompatibility. The grasp that affords reliable actuation can constrain the hand so that the subsequent target pose is unreachable without collision. For example, illuminating the toy requires a button-side grasp to press the switch, yet that orientation forces the hand to collide with the table when placing the light. Resolving such conflicts would require capabilities such as in-hand reorientation or finger gaiting.

Failures also occurred during in-hand actuation, revealing sensitivity to precise contact physics and sim2real gaps, particularly evident in these three challenging tasks. Both `clean keyboard` and `illuminate toy` task demand high precision on button pressing—especially with the flashlight's small ($< 1cm$), soft button; slight contact misalignments frequently caused the finger to slip off before successful actuation. The `grind salt` task primarily suffered from a sim2real gap, manifesting as slippage and grinder toppling. Slippage can likely be reduced with more aggressive friction domain randomization (and improved surface modeling). Toppling appears to stem from mismatches in controller dynamics and contact modeling between simulation and hardware; mitigating it will require careful controller gain calibration and contact/collision parameter tuning.

These challenging cases indicate that while our method significantly advances CD-FOM, robustly handling scenarios demanding extreme precision (like `illuminate toy`) or subject to substantial sim2real variations in contact physics (like `grind salt`) remains an important direction for future work, potentially involving richer object modeling and learning adaptive policies.

Q2 — Quality of VLM-CEM Generated Constraints

While we use binary criteria in assessing whether an object final pose is correct in Q1, some of the final poses are more better than the others (e.g. a spray bottle that aims at the edge of the plant versus aiming at the center of the plant). To evaluate the quality of our novel VLM-CEM procedure for generating global constraints, we asked twenty participants to rate the rendered images generated based on the resulting goal poses from the VLM-CEM and three additional baselines:

- **VLM-CEM**: our keypoint-anchored sampler that generates candidate goal poses around detected interaction points on the object.
- **VLM-CEM (Dir.)**: a variant that restricts translations sampling to be along the object's functional axis (e.g., nozzle direction).
- **PIVOT (SE3)** [30]: an adaptation of PIVOT that perturbs full 6-DoF poses in image-space without explicit keypoint anchoring, often resulting in misalignment in depth or lateral offset.
- **PIVOT (Trans.)** [30]: akin to the original PIVOT method, searching only in 2D image-space translations based on coarse visual alignment.

We generate three multi-view images per task per method and requested human ratings on a five-point scale (1 = unreasonable, 5 = perfect), where a human rating of (≥ 3) is considered as semantically acceptable in our analysis.

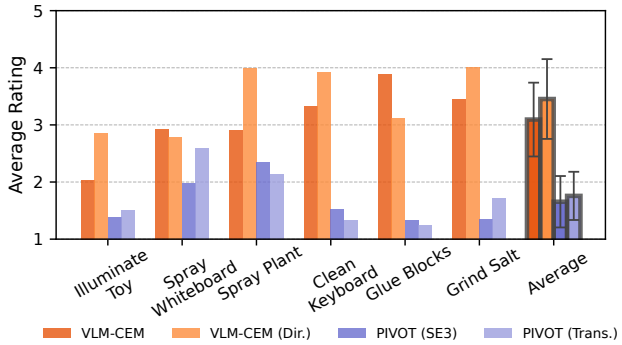

Fig. 8: Human study ratings of generated global goals. We request human feedback on the global goals generated by our VLM-CEM procedure and baselines (VLM-CEM without rotation changes, PIVOT with rotation and without rotations). We also report the average and standard deviation error bars of the results across all goals for each respective method. On average, the two VLM-CEM methods (ours) are ranked higher in most tasks.

Fig. 9: Example visualizations of different goal-pose-generation methods on the task `clean keyboard`. Both variants of VLM-CEM generate **both semantically and physically valid** global constraints, while the baseline methods perform poorly on the task.

Fig. 8, depicts the result of our human ratings. We observe that our VLM-CEM, along with the variant with directional exploration, is ranked highest by humans in all tasks. To ensure the validity of this finding, we perform a Wilcoxon Signed-Rank test and verify this result to be significant with a $p < 0.02$ for each comparison between the 2 PIVOT methods and our 2 VLM-CEM methods.

Surprisingly, for the `spray plant` task, our VLM-CEM with directional exploration ablation scores significantly higher than the VLM-CEM with full translational exploration. We hypothesize this is because the directional constraint, while reducing the exploration space, effectively filters out candidate poses that might appear plausible in the 2D rendered images used for VLM scoring but are functionally misaligned in 3D (e.g., aiming near the plant but slightly off-axis). By enforcing alignment along the functionally critical nozzle direction, the directional variant ensures better geometric task relevance for the highest-scoring poses in this specific spraying task.

By manually inspecting the lowest-scoring cases, we observe that most of them correspond to PIVOT's results and they are caused because they rely purely on image-space alignment and thus often propose poses that appear correct in 2D yet place the object off the interaction line in 3D (e.g., a flashlight "aiming" at the toy but missing it laterally). In contrast, VLM–CEM samples poses anchored at detected interaction keypoints, yielding goals that are both visually

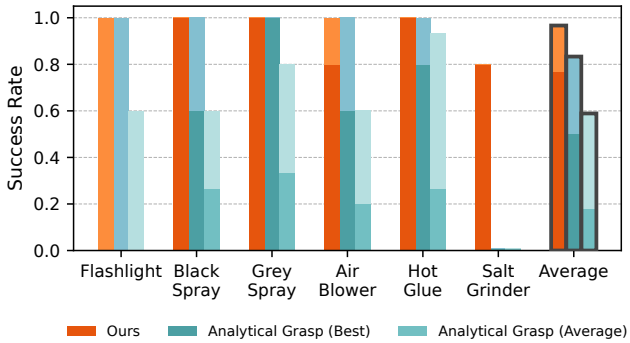

Fig. 10: Performance gains of CoDex constraint-guided policy training compared to the direct execution of the 3 and the best analytical grasps from CoDex's constrained optimization. Total bar height indicates the success rate of achieving a stable grasp through lifting. The bottom segment (darker shade) represents the success rate of achieving both a stable grasp *and* successful actuation. By training with constraint-guided RL in simulation for the full task, CoDex significantly improves stability and actuation of the objects.

and spatially reasonable.

Q3 — Benefits of Constraint-Guided RL Policy Learning

In this set of experiments, we evaluate the improvements gained from our compositional policy learning stage by comparing its **stability** against executing the **analytical** grasp generated via constrained optimization (adapted from [23]). We compare our learned policy (CoDex) against the direct execution of these analytical grasps (evaluating 3 grasps per object, each repeated 5 times with random object initialization). We report comparisons against both the average performance across all grasp candidates per object and the *oracle best performance*—the success rate achieved by the single most successful grasp candidate for each object, selected post-hoc after evaluating all grasp candidates. Fig. 10 depicts the results.

We observe that our constraint-guided policy learning significantly improves grasp stability, exceeding the average performance of initial grasp candidate by over 36% and surpassing the *oracle best performance* (the maximum potential success achievable without refinement) by over 12%. The benefit of policy learning is even more pronounced for functional actuation: the learned policy achieves 60% higher actuation success than the average grasp candidate and crucially, over 26% higher success than the best possible outcome using only the initial grasp candidates (oracle).

Interestingly, none of the initial grasp candidates achieved stable grasping success (let alone actuation) for the challenging salt grinder, likely due to its slippery surface and geometry. However, CoDex's policy learning stage successfully discovers a stable grasp, highlighting the method's ability to improve even on difficult cases. This demonstrates the importance of the holistic, simulation-based policy learning stage. It allows CoDex to refine statically plausible grasp candidates into dynamically robust policies that significantly enhance functional viability compared to executing the initial grasp candidates directly, even when considering the best possible initial grasp candidate.

V. Limitations

While CoDex demonstrates success across multiple compositional FOM tasks, it also reveals several limitations that we plan to address in future work. First, tasks requiring pinpoint contact, such as pressing tiny push-buttons, are sensitive to finger size and actuator tolerances, demanding higher accuracy in control and possibly different hand morphology to ensure reliable execution. Second, many tasks go beyond reaching a single goal pose and instead require sustained, coupled arm–hand motion—for example, actuating scissors while sliding along paper—calling for extensions toward trajectory-level constraints and closed-loop feedback. Finally, the current policy assumes a single actuation point, leaving out objects that need alternating or multi-point actuation; extending CoDex to these tasks will open new robot capabilities.

VI. Conclusion

We addressed the challenging problem of zero-demonstration functional object manipulation by introducing CoDex, a framework that translates abstract VLM guidance into concrete semantic constraints. Our method enforces these VLM-generated local (e.g., actuation points) and global (e.g., target poses) constraints through a two-phase process: first, analytic constrained optimization efficiently generates a set of stable, function-aligned grasp candidates, and second, constraint-guided reinforcement learning initializes from these candidates to discover the complete, dynamically robust grasp-move-actuate policy. Our experiments on a physical robot demonstrated that this tight integration of semantic reasoning with a physics-grounded, constraint-enforcing pipeline is crucial. CoDex autonomously discovered and executed complex strategies for diverse tasks with unseen objects, validating our approach. This work represents a key step toward versatile and autonomous tool manipulation. Future directions include extending the framework to more complex object mechanisms, incorporating tactile feedback for finer control, and exploring richer, interactive VLM dialogues.

References

[1] D. Paulius, Y. Huang, R. Milton, W. D. Buchanan, J. Sam, and Y. Sun, "Functional object-oriented network for manipulation learning," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 2655–2662.

[2] K. Srinivasan, E. Heiden, I. Ng, J. Bohg, and A. Garg, "Dexmots: Learning contact-rich dexterous manipulation in an object-centric task space with differentiable simulation," in *International Symposium on Robotics Research (ISRR)*, 2024.

[3] L. Huang, H. Zhang, Z. Wu, S. Christen, and J. Song, "Fungrasp: Functional grasping for diverse dexterous hands," *IEEE Robotics and Automation Letters*, 2025.

[4] M. Aburub, K. Higashi, W. Wan, and K. Harada, "Functional eigengrasping using approach heatmaps," *arXiv preprint*, 2024.

[5] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," in *Conference on Robot Learning (CoRL)*, 2023.

[6] M. Li, Z. Chen, C. Yang, and Q. Zhu, "Dexterous manipulation with multi-fingered robotic hands: A review," *Frontiers in Neurorobotics*, vol. 16, p. 861825, 2022.

[7] S. An, Z. Meng, C. Tang, Y. Zhou, T. Liu, F. Ding, S. Zhang, Y. Mu, R. Song, W. Zhang, Z.-G. Hou, and H. Zhang, "Dexterous manipulation through imitation learning: A survey," *arXiv preprint arXiv:2504.03515*, 2025.

[8] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[9] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*, 2022, pp. 1678–1690.

[10] A. Iyer, Z. Peng, Y. Dai, I. Guzey, S. Haldar, S. Chintala, and L. Pinto, "Open teach: A versatile teleoperation system for robotic manipulation," *arXiv preprint arXiv:2403.07870*, 2024.

[11] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[12] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[13] V. Kumar, A. Gupta, E. Todorov, and S. Levine, "Learning dexterous manipulation policies from experience and imitation," *arXiv preprint arXiv:1611.05095*, 2016.

[14] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 778–13 790.

[15] A. Bahety, P. Mandikal, B. Abbatematteo, and R. Martín-Martín, "Screwmimic: Bimanual imitation from human videos with screw space projection," in *Robotics: Science and Systems*, 2024.

[16] A. Bahety, A. Balaji, B. Abbatematteo, and R. Martín-Martín, "Safemimic: Towards safe and autonomous human-to-robot imitation for mobile manipulation," in *Robotics: Science and Systems*, 2025.

[17] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[18] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[19] K. B. Shimoga, "Robot grasp synthesis algorithms: A survey," *The International Journal of Robotics Research*, vol. 15, no. 3, pp. 230–266, 1996.

[20] A. T. Miller and P. K. Allen, "Graspit!: A versatile simulator for grasp analysis," in *ASME International Mechanical Engineering Congress and Exposition*, vol. 26652. American Society of Mechanical Engineers, 2000, pp. 1251–1258.

[21] D. Berenson and S. S. Srinivasa, "Grasp synthesis in cluttered environments for dexterous hands," in *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2008.

[22] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[23] A. H. Li, P. Culbertson, J. W. Burdick, and A. D. Ames, "Frogger: Fast robust grasp generation via the min-weight metric," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6809–6816.

[24] H. Charlesworth and G. Montana, "Solving challenging dexterous manipulation tasks with trajectory optimisation and reinforcement learning," in *Proceedings of the 3rd Workshop on Machine Learning for Autonomous Driving, PMLR*, vol. 139, 2021.

[25] S. Chen, J. Bohg, and C. K. Liu, "Springgrasp: Synthesizing compliant, dexterous grasps under shape uncertainty," *arXiv preprint arXiv:2404.13532*, 2024.

[26] J. Zhang, W. Xu, Z. Yu, P. Xie, T. Tang, and C. Lu, "DexTOG: Learning Task-Oriented Dexterous Grasp with Language Condition," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, 2025.

[27] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[28] J. Wen, Y. Zhu, J. Li, Z. Tang, C. Shen, and F. Feng, "Dexvla: Vision-language model with plug-in diffusion expert for general robot control," *arXiv preprint arXiv:2502.05855*, 2024.

[29] J. Liu, M. Liu, Z. Wang, P. An, X. Li, K. Zhou, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation," *arXiv preprint arXiv:2406.04339*, 2024.

[30] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, Q. Vuong, T. Zhang, T.-W. E. Lee, K.-H. Lee, P. Xu, S. Kirmani, Y. Zhu, A. Zeng, K. Hausman, N. Heess, C. Finn, S. Levine, and B. Ichter, "Pivot: Iterative visual prompting elicits actionable knowledge for vlms," 2024.

[31] W. Huang, C. Wang, Y. Li, R. Zhang, and F.-F. Li, "Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation," in *Conference on Robot Learning (CoRL)*, 2024.

[32] H. Liu, S. Guo, P. Mai, J. Cao, H. Li, and J. Ma, "Robodexvlm: Visual language model-enabled task planning and motion control for dexterous robot manipulation," *arXiv preprint*, 2025.

[33] Z. Li, J. Liu, Z. Li, Z. Dong, T. Teng, Y. Ou, D. Caldwell, and F. Chen, "Language-guided dexterous functional grasping by llm generated grasp functionality and synergy for humanoid manipulation," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 10 506–10 519, 2025.

[34] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 6396–6403.

[35] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4269–4276.

[36] C. Ferrari and J. F. Canny, "Planning optimal grasps," in *Proceedings., IEEE International Conference on Robotics and Automation*. IEEE, 1992, pp. 2290–2295.

[37] L. Wang, Y. Xiang, and D. Fox, "Manipulation trajectory optimization with online grasp synthesis and selection," in *Robotics: Science and Systems (RSS)*, 2020.

[38] T. Weng, D. Held, F. Meier, and M. Mukadam, "Neural grasp distance fields for robot manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[39] B. Sundaralingam, A. Lambert, C. Wang, Y. Li, F.-F. Li, and R. Zhang, "Multi-finger manipulation via trajectory optimization with differentiable rolling and geometric constraints," *arXiv preprint arXiv:2408.13229*, 2024.

[40] B. Zhou, H. Yuan, Y. Fu, and Z. Lu, "Learning diverse bimanual dexterous manipulation skills from human demonstrations," *arXiv preprint arXiv:2410.02477*, 2024.

[41] P. Koczy, M. C. Welle, and D. Kragic, "Learning dexterous in-hand manipulation with multifingered hands via visuomotor diffusion," *arXiv preprint arXiv:2503.02587*, 2025.

[42] C. Wang, R. Yang, J. Ichnowski, M. Danielczuk, Z. Xian, C. Gonzalez, R. H. Taylor, K. Goldberg, P. Abbeel, C. H. Rycroft, and Y. Ma, "Kinesoft: Learning proprioceptive manipulation policies with soft robot hands," *arXiv preprint arXiv:2503.01078*, 2025.

[43] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[44] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[45] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, J. Lu, T. Anderson, E. Bransom, K. Ehsani, H. Ngo, Y. Chen, A. Patel, M. Yatskar, C. Callison-Burch, A. Head, R. Hendrix, F. Bastani, E. VanderBilt, N. Lambert, Y. Chou, A. Chheda, J. Sparks, S. Skjonsberg, M. Schmitz, A. Sarnat, B. Bischoff, P. Walsh, C. Newell, P. Wolters, T. Gupta, K.-H. Zeng, J. Borchardt, D. Groeneveld, C. Nam, S. Lebrecht, C. Wittlif, C. Schoenick, O. Michel, R. Krishna, L. Weihs, N. A. Smith, H. Hajishirzi, R. Girshick, A. Farhadi, and A. Kembhavi, "Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models," 2024.

[46] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "Triposr: Fast 3d object reconstruction from a single image," *arXiv preprint*, 2024.

[47] S. Tao, F. Xiang, A. Shukla, Y. Qin, X. Hinrichsen, X. Yuan, C. Bao, X. Lin, Y. Liu, T. kai Chan, Y. Gao, X. Li, T. Mu, N. Xiao, A. Gurha, V. N. Rajesh, Y. W. Choi, Y.-R. Chen, Z. Huang, R. Calandra, R. Chen, S. Luo, and H. Su, "Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai," 2025.