ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

VLG Computer Vision and Learning Group

ScrubGrasp: Grasp and Affordance Synthesis for Surgical Instruments via Scalable Digital Twins

Master's Thesis

Ankita Ghosh
Department of Computer Science

External Supervisor
Dr. Lalithkumar Seenivasan
Prof. Dr. Mathias Unberath
Department of Computer Science
Johns Hopkins University
USA

Internal Supervisor
Frano Rajič
Prof. Dr. Siyu Tang
Department of Computer Science
ETH Zurich
Switzerland

May 31, 2026

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

ScrubGrasp: Grasp and Affordance Synthesis for Surgical Instruments via Scalable Digital Twins

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Ghosh	Ankita

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zurich, 31 May 2026	

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).

Abstract

Robust grasping of surgical instruments is critical for clinical automation, yet training robust grasping policies is hindered by the scarcity of physics-verified, context-aware, annotated data. We introduce ScrubGrasp, a zero-shot synthesis framework that serves as an automated grasp affordance annotation engine. From a single RGB-D input, ScrubGrasp employs vision foundation models to build scene-aware digital twins, autonomously generating 2D and 3D grasp affordance on simulated, physically stable grasps. Evaluated on a custom-curated benchmark of 120 cluttered scenes containing 16 clinical tools across 600+ poses, ScrubGrasp outperforms state-of-the-art grasp synthesis baselines by 2.2x. We validate the framework's utility by bootstrapping a downstream affordance policy trained entirely on affordance annotations generated by ScrubGrasp. The learned model effectively targets unoccluded graspable regions, demonstrating ScrubGrasp's capability to provide the scalable data infrastructure required for training end-to-end vision models in safety-critical environments.

Contents

1 Introduction **1**

2 Related Work **4**
2.1 Single-Object Dexterous Grasp Synthesis 4
2.2 Dexterous Grasp Synthesis in Clutter 4
2.3 Affordance-Guided Robotic Manipulation 5
2.4 Robotic Handling of Surgical Instruments 5
2.5 Digital Twins and Foundation Models in Surgery 6

3 Preliminary **7**
3.1 Lightning Grasp 7
3.2 Segment Anything Model 3 7
3.3 Segment Anything Model 3D Objects 8
3.4 Isaac Lab 8

4 Methodology **9**
4.1 Overview: ScrubGrasp 9
4.2 Digital Twin Representation 9
4.2.1 Instrument Digital Twin 9
4.2.2 Scene Alignment 10
4.3 Grasp Synthesis 11
4.3.1 Framework 11
4.3.2 Conditioning Modules 11
4.4 Affordance Generation 12
4.4.1 3D Grasp Affordance 12
4.4.2 Grasp Candidate Stability Validation 13
4.4.3 2D Grasp Affordance 14
4.5 Automated Scene Affordance Annotation 14

5 Experiments and Results **15**
5.1 Experimental Setup 15
5.2 Surgical Instrument Dataset 15
5.3 Evaluation Metrics 16

5.4 Simulation-Based Grasp Validation . . . 17
5.5 Comparison against State-of-the-Art Grasp Synthesis Methods . . . 18
5.6 Annotation Validation via Downstream Affordance Prediction . . . 20
5.7 Ablation Studies . . . 21
5.7.1 Ablation Study on Conditioning modules . . . 21
5.7.2 Digital Twin Representation . . . 22

6 Discussion **23**
6.1 Applications . . . 23
6.2 Limitations . . . 24
6.3 Future Work . . . 24

7 Conclusion **26**

A Supplementary Material **28**

List of Figures

1.1 Grasp and affordance synthesis for diverse surgical instruments. 2

4.1 Overall methodology: (A) *ScrubGrasp*: (a) The pipeline takes an RGB-D image and uses SAM 3 and SAM 3D to generate the surgical tool meshes. It then performs scene alignment using SVD-based registration. (b) It conditions our backbone grasp generation model, modified Lightning Grasp, on bounded sampling (ϕ_{BS}), grasp surface conditioning (ϕ_{GS}), and scene collision conditioning (ϕ_{SC}) modules to produce grasp candidates. (c) These are then filtered via a physics engine to compute 3D and 2D grasp affordance using geodesic mesh traversal and backprojection, respectively. (B) This framework iteratively aggregates instrument affordances from RGB-D inputs into a scene-level annotation. 10

5.1 Sample scenes from our custom-curated dataset across six clutter configurations: loose (easy, hard), dense (easy, hard), and overlap (easy, hard). Easy scenes contain larger instruments with accessible, graspable regions, while hard scenes include thin, delicate instruments with complex geometries. 16

5.2 Qualitative results across six clutter configurations. For each target instrument: digital twin, synthesized grasp, contact heatmap, and 2D affordance. Digital twin: green (graspable regions), red (non-graspable due to occlusions, overlap, or surface contact), blue (neighboring obstacle point cloud). Contact and affordance heatmap: red (full graspability) to blue (≈ 0). 19

5.3 Qualitative comparison of ScrubGrasp affordance annotations vs. downstream model predictions on our custom dataset. Heatmap: red (full graspability) to blue (≈ 0). 21

5.4 Ablation study on SAM 3 inference time in generating instrument digital twin based on total number of tools in the scene. 22

List of Tables

5.1 Grasp success rates (%) compared with state-of-the-art methods across six clutter configurations (loose, dense, overlap × easy, hard) and robotic hands with different joint degrees of freedom (DOF). 18
5.2 Comparison of key method properties. 20
5.3 Affordance prediction model performance on test set (20 scenes). 21
5.4 Ablation study evaluating the contribution of each conditioning module. LG: Lightning Grasp [65] (base grasp generation), BS: bounded sampling, GS: graspable surface conditioning, SC: scene collision conditioning. SR: success rate (%), TC: time cost (seconds). . . 22

Chapter 1

Introduction

Enabling robots to *reliably grasp and handle* surgical instruments offers transformative benefits across the perioperative workflow – spanning pre-operative tray preparation, intra-operative surgical assistance, and post-operative reprocessing – with significant *clinical* and *workforce* implications. *Clinically*, it can help mitigate risks to patient safety from errors in manual sorting and counting. Inconsistent manual sorting in the sterile processing department could lead to incomplete cleaning or improper tray assembly, risking patient contamination [44, 40]. Errors in manual counting during surgery are equally problematic: retained surgical instruments – a serious "never event" – occur in approximately 1 in 5,000-10,000 procedures [13, 48], with manual counting remaining error-prone, especially due to staff fatigue during prolonged cases lasting 2-12+ hours [2, 58]. Reliable robotic grasping enables the deployment of robotic systems that enforce consistent, traceable handling – logging every grasp event to support instrument tracking and automated counting [12] – while also simplifying sterilization through standardized orientation and spacing that optimize sterilant penetration.

From a *workforce* perspective, reliable grasping unlocks the use of robotic systems in addressing a pressing personnel crisis. Scrub nurses, surgical technologists, and sterile processing technicians face vacancy rates exceeding 15% worldwide [57], with training requiring 6-24 months of institutional investment. Robots capable of effective instrument handling can take over repetitive tasks, easing the burden on scarce staff, reducing fatigue-induced errors, and limiting exposure to sharp injuries that affect hundreds of thousands of workers annually [17, 41, 6].

Despite this clear need, scalable robotic grasping of surgical instruments remains an open challenge. Tools exhibit extreme geometric diversity – from slender needle holders to bulky retractors – and appear in diverse clutter configurations, such as scattered, compactly packed, or overlapping arrangements on trays [39]. Safe manipulation in risk-averse healthcare settings demands grasps that are not merely physically stable but context-aware and functionally safe, highlighting the importance of "*graspable affordance*": defining *where* an instrument must be grasped safely in the given scene. Grasps generated purely from geometric stability often become infeasible when they ignore semantic boundaries and scene clutter. Yet, training policies to recognize these functional affordances is bottlenecked by the prohibitive cost of annotating physics-verified, semantically safe data – further exacerbated by privacy, sterility, and workflow constraints in healthcare [1]. Existing data-driven approaches [31, 69, 63, 27] consequently struggle to generalize across diverse surgical specialties without extensive retraining.

Existing grasp synthesis methods remain limited by their dependence on training data, CAD models,

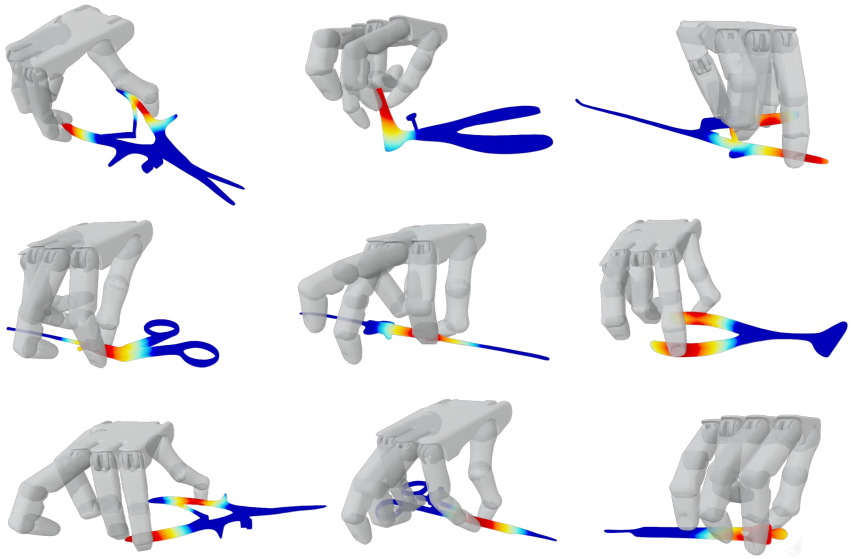


Figure 1.1: Grasp and affordance synthesis for diverse surgical instruments.

or object-centric generation that does not account for scene-level constraints – limiting generalizability and scalability to diverse instruments and scenes. Learning-based approaches such as DDGC [31], DexGraspNet 2.0 [69], UniDexGrasp [63], and HGC-Net [27] achieve impressive results on benchmark objects, but may require retraining to adapt to new surgical instruments or scenes – limiting their applicability in clinical settings where tool inventories evolve frequently. CAD-based pipelines such as GraspIt! [33] depend on accurate 3D models for each instrument, which may not be readily available and impose significant maintenance burdens as catalogs change. Optimization-based methods like Lightning Grasp [65] can operate on reconstructed meshes without requiring CAD models. However, they typically generate grasps in free space without scene-level awareness, producing candidates that may collide with surrounding objects or supporting surfaces. There is a need for a zero-shot, scene-aware grasp synthesis pipeline that can generalize across diverse surgical instruments and scene arrangements – without requiring training data, CAD models, or instrument-specific programming.

To overcome these challenges, we introduce ScrubGrasp, a training-free and CAD-model-free zero-shot grasp and affordance synthesis framework (Figure 1.1) that generates physically plausible, scene- and geometry-conditioned grasps for surgical instruments from a single RGB-D image. Our key contributions are as follows:

- A zero-shot grasp and affordance synthesis framework that leverages vision foundation models to construct scene-aware digital twins from a single RGB-D image. By capturing the full geometric context of both the target tool and surrounding clutter without requiring explicit CAD models, the framework autonomously generates high-quality 2D and 3D grasp affordances.

- An automated grasp affordance annotation engine built upon simulation-verified, physically stable grasps. We validate the utility of this scalable infrastructure by successfully bootstrapping a down-

stream affordance policy— trained entirely on ScrubGrasp-generated annotations— that effectively targets unoccluded graspable regions for safe, end-to-end manipulation in dynamic clinical settings.

- A custom-curated surgical instrument benchmark designed to systematically evaluate surgical grasp synthesis, featuring 120 cluttered scenes, 16 distinct clinical tools, and over 600 poses across various clutter configurations.

The remainder of this report is organized as follows. Chapter 2 reviews related work concerning dexterous grasp synthesis, affordance-guided robotic manipulation, and the integration of digital twins and foundation models in surgery. Chapter 3 introduces the preliminary frameworks and models essential to our approach, including Lightning Grasp, the Segment Anything Models (3 and 3D Objects), and Isaac Lab. In Chapter 4, we detail the proposed methodology, providing an overview of the ScrubGrasp framework and elaborating on the digital twin representation, grasp synthesis pipelines, and automated affordance generation. Chapter 5 presents the experiments and results. It describes the experimental setup and surgical instrument dataset before detailing comprehensive evaluations, which include simulation-based validations, comparisons against state-of-the-art methods, and ablation studies. Finally, Chapter 6 discusses potential applications, current limitations, and avenues for future work, followed by concluding remarks in Chapter 7.

Chapter 2

Related Work

2.1 Single-Object Dexterous Grasp Synthesis

The foundational challenge in dexterous robotic manipulation has historically centered on synthesizing stable grasps for isolated objects. Early approaches primarily relied on analytical methods, utilizing precise geometric models to compute optimal contact points, minimize wrench spaces, and ensure force closure [11, 33]. While mathematically rigorous, these techniques often struggle to generalize across novel object geometries and require expensive real-time computation for multi-fingered hands with higher degrees of freedom. To address these bottlenecks, the introduction of fast, differentiable force closure estimators [28] catalyzed a paradigm shift toward data-driven, simulation-based generative models by enabling the rapid synthesis of massive training datasets.

Capitalizing on this accelerated estimation, DexGraspNet [54] provides millions of high-quality, physics-validated dexterous grasps across diverse object geometries. By training on such large-scale synthetic datasets– often alongside human-interaction captures like ContactDB [3] and DexYCB [4]– deep learning frameworks can now predict complex joint configurations directly from implicit visual representations [63, 25]. Expanding on these learning-based techniques, frameworks such as OmniDexVLG [72] incorporate vision-language guidance to generate grasping strategies, while methods like Dex1B [64] and sim-to-real reinforcement learning pipelines [30, 47] rely on vast amounts of demonstration data to bridge the embodiment gap. Parallel to these data-driven advances, modern optimization-based planners offer robust, training-free alternatives by focusing heavily on contact modeling and geometric stability. For instance, recent methods like OmniDexGrasp [56] and Lightning Grasp [65] utilize force feedback and contact fields to procedurally synthesize high-performance grasps.

Despite these advancements, single-object methods assume an unobstructed target with clear boundaries. Consequently, their performance degrades significantly in unconstrained, densely populated environments, necessitating a shift toward context-aware strategies capable of navigating scene-level physical constraints.

2.2 Dexterous Grasp Synthesis in Clutter

Early generative models like DDGC [31] and single-shot networks like HGC-Net [27] pioneered multi-finger grasp prediction by explicitly encoding scene information to navigate obstacles. Modern learning-based approaches have demonstrated significant generalizability by leveraging massive synthetic environments.

Models like DexGraspNet 2.0 [69] and DexGrasp Anything [73] integrate explicit physical constraints into diffusion-based generation, ensuring geometrically feasible and robust grasping poses across an immensely diverse set of novel objects. ClutterDexGrasp [5] has developed advanced zero-shot sim-to-real systems to handle severe occlusions, utilizing teacher-student distillation for dynamic target-oriented grasping. To explicitly mitigate complex inter-object collisions in dense piles, models like CADGrasp [70] integrate contact-aware representations directly into the optimization pipeline.

While these methods excel in general environments, their reliance on massive training datasets poses a significant challenge in the highly specialized healthcare domain. Extensive, domain-specific annotated datasets for surgical environments are scarce and costly to produce, severely limiting the adaptability of these models to evolving clinical tool inventories. Moreover, even when a synthesized grasp successfully navigates clutter and achieves geometric stability, it inherently lacks the functional context required to ensure a surgical tool is held safely and correctly for downstream application.

2.3 Affordance-Guided Robotic Manipulation

To bridge this critical gap between physical stability and functional utility, manipulation research has increasingly integrated explicit affordance synthesis to define interaction semantics. Early foundational frameworks, such as AffordanceNet [10], formulated affordance detection as a dense pixel-level segmentation task. Subsequent methods like Where2Act [35] shifted this paradigm toward learning actionable 3D visual priors from simulated interactions. More recently, the field has leveraged visual foundation models to enable one-shot, open-vocabulary affordance learning [22], drastically reducing the reliance on exhaustive, category-specific training sets.

Building upon these synthesis methods, recent affordance-guided architectures explicitly encode these learned visual priors to execute task-oriented manipulation in complex environments. For example, multi-affordance matching frameworks have successfully enabled stable pick-and-place execution for novel objects in dense clutter [66]. To extend these capabilities to higher-dexterity multi-fingered effectors, models such as ContactDexNet [71] integrate dense hand-object contact semantic mapping. In a parallel approach, other methods extract precise, human-like interaction priors directly from egocentric video demonstrations [23]. Pushing toward generalized, high-level reasoning, cutting-edge systems now employ large-scale affordance segmentation [60] and in-context learning [49] to deduce task-specific grasping strategies for open-vocabulary targets dynamically. Motivated by these advancements, our work centers on *grasp affordances* because in the high-stakes healthcare domain, merely securing a target is insufficient. The system must know specifically *where* to grasp a surgical tool to ensure functionally viable clinical execution.

2.4 Robotic Handling of Surgical Instruments

Translating grasping techniques to the surgical domain introduces strict constraints regarding safety, precision, and specialized tool geometry. Early automation relied on single-view vision algorithms and electromagnetic grippers to extract isolated tools from cluttered trays [62]. Recent advances have extended this to sterile processing, employing deep object detection and rule-based packing to automate tray assembly while minimizing collisions [43]. For dynamic, real-time surgical assistance, modern robotic scrub nurses increasingly leverage multimodal reasoning. Systems like RoboNurse-VLA [26] integrate vision-language-action

models to execute precise, voice-controlled instrument handovers.

In these high-stakes settings, respecting grasp affordance is a strict safety requirement rather than a mere optimization metric [61, 32]. However, deploying affordance-led models is severely bottlenecked by the prohibitive cost of annotating physics-verified, semantically safe data. Existing pipelines relying on exhaustive human labeling [18] or engineered geometric features [36, 55] simply cannot scale to the extreme geometric diversity of clinical instruments.

2.5 Digital Twins and Foundation Models in Surgery

To overcome the limitations of manual annotation and physical data collection, digital twins are increasingly being adopted across robotics for scalable simulation, data generation, and policy learning [16, 24, 9]. Within the operating room (OR), digital twins have been rapidly adapted for a variety of tasks, including surgical simulation and spatial planning [15, 20], privacy-preserving workflow analysis [37], and the photorealistic rendering of dynamic OR environments [68, 67].

Concurrently, the emergence of vision foundation models has fundamentally transformed the pipeline for scene perception and reconstruction. At the 2D level, the Segment Anything Model (SAM) [19] introduced unprecedented zero-shot generalization for object segmentation across diverse visual domains. This foundational semantic understanding is further enriched by advanced self-supervised vision transformers, such as DINOv3 [46], which extract dense, highly discriminative feature representations. To elevate these 2D visual priors into the spatial domain, recent architectures like the Visual Geometry Grounded Transformer (VGGT) [53] explicitly integrate 3D geometric reasoning directly into the visual feature extraction process. Building upon these intersecting capabilities, specialized frameworks such as SAM 3D [42] seamlessly bridge the gap between pixel-level segmentation and spatial generation, enabling robust, zero-shot 3D reconstruction directly from standard imaging. Building on these capabilities, our pipeline leverages foundation models to construct scalable, scene-aware digital twins from a single RGB-D image. We utilize this infrastructure not only to validate grasp stability through physics-based simulation but also to serve as an automated engine for computing grasp affordances.

Chapter 3

Preliminary

3.1 Lightning Grasp

As a preliminary foundation for our proposed framework, we adopt the Lightning Grasp [65] algorithm as the primary backbone for grasp generation. Lightning Grasp is an analytical procedural synthesis algorithm that rapidly computes valid grasps for dexterous hands by introducing a novel "contact field" concept. The contact field is a 6D geometry object that characterizes spatial contacts that a hand mesh can potentially generate, encoding both position and normal. This field is structured as a Bounding Volume Hierarchy (BVH) [59], enabling the algorithm to verify valid hand-object contacts by casting rays outward from the object's surface to detect collisions with the kinematic links of the hand. This abstraction reduces contact domain detection to a collision problem and decouples complex geometric computations from the optimization search.

A significant advantage of this formulation is its training-free efficiency. The algorithm is capable of generating a diverse set of grasps in a matter of seconds without relying on prior training datasets, manual hand initialisation templates, or sensitive parameter tuning. Despite these computational benefits, the baseline algorithm possesses inherent limitations that preclude its direct deployment in unstructured environments. It relies heavily on highly accurate 3D perception, requiring pristine object meshes or high-quality point clouds to reliably synthesize contacts. Furthermore, Lightning Grasp lacks scene-level spatial awareness, operating under the assumption that target objects exist in a vacuum. As it does not natively account for surrounding environmental clutter, table surfaces, or tight operational spaces, leveraging its high-performance synthesis in real-world scenarios necessitates the conditioning modules introduced in our proposed method.

3.2 Segment Anything Model 3

To enable digital twin generation of our surgical instruments from RGB-D images, we adopt the Segment Anything Model 3 (SAM 3) [19] as the backbone for semantic segmentation. SAM 3 is a unified vision foundation model that performs promptable "concept segmentation", extracting and tracking all instances of a target class based on open-vocabulary text or visual exemplars. SAM 3 uses a transformer-based [52] detector that combines a high-capacity vision encoder and a text encoder into a unified representation. By processing prompt tokens and visual data simultaneously, the model achieves generalized deep semantic understanding.

However, SAM 3 is fundamentally confined to the 2D image domain, relying entirely on pixel space without any intrinsic depth awareness or 3D spatial reasoning. To bridge the gap between 2D semantic masks and the 3D operational workspace, we couple it with SAM 3D Objects in our proposed method.

3.3 Segment Anything Model 3D Objects

We utilize the SAM 3D Objects [42] foundation model for single-view geometric recovery. SAM 3D Objects is a generative vision model that reconstructs full 3D shape geometry, texture, and spatial layout directly from a single 2D image. This architecture employs a multi-stage Diffusion Transformer framework [7] that first infers the 3D object's shape and canonical pose from an input mask– provided by SAM 3 in our pipeline– and then refines the fine-grained texture, effectively lifting flat semantic segments into volumetric assets.

A significant advantage of this model is its zero-shot robustness in uncurated environments. It is capable of generating independent, posed 3D assets of complex targets, such as surgical instruments, even in the presence of occlusion and visual clutter. Consequently, we leverage this powerful 3D generation, augmented by our own geometric refinement and scene alignment, to construct the digital twin representations that serve as the elementary step of our grasp and affordance synthesis pipeline.

3.4 Isaac Lab

We adopt the NVIDIA Isaac Lab [34] framework for the simulation and physical verification of grasp stability. Isaac Lab is a GPU-accelerated, modular robot learning framework built upon Omniverse and Isaac Sim that enables high-fidelity physics and realistic sensor simulation at data-center scale. This architecture leverages massive parallelization across thousands of simultaneous environments, tightly coupling PhysX-based multi-body dynamics with RTX rendering to evaluate complex kinematic interactions entirely within the GPU. By natively supporting complex embodiments, such as dexterous robotic hands, and accurately computing their associated contact dynamics, Isaac Lab facilitates the rapid and scalable validation of synthesized grasping behaviors.

Chapter 4

Methodology

4.1 Overview: ScrubGrasp

We propose *ScrubGrasp* (Figure 4.1), which takes as input (a) an RGB-D image $\mathcal{I}$ of a scene containing N instruments $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_N\}$, and (b) a robot hand model $\mathcal{R}$. to generate a set of scene-aware, physics-validated grasps $\mathcal{G} = \{g_k\}_{k=1}^{K}$ and a corresponding 2D grasp affordance mask $\mathcal{A}$ that overlays on I for a target instrument $\tau_n \in \mathcal{T}$. *ScrubGrasp* constructs a scalable digital twin of the instrument and scene from $\mathcal{I}$ at inference time, bypassing the need for predefined CAD models. This Digital twin representation facilitates contextual conditioning $\phi(\cdot)$, extending Lightning Grasp [65] to generate grasp candidates conditioned on geometry-aware plausible global hand poses, graspable surface regions, and scene-level collision constraints. The candidate grasps are evaluated for physical stability via the simulation engine ρ, aggregated into a 3D grasp affordance, and subsequently backprojected to derive the final 2D affordance mask. Formally, *ScrubGrasp* is defined as:

$$\left(\mathcal{A}_{\tau_n}, \mathcal{G}_{\tau_n}\right) = \text{ScrubGrasp}_{\phi(\tau_n),\rho}(\mathcal{I}, \mathcal{R}) \tag{4.1}$$

where $\phi(\tau_n)$ denotes the scene-aware conditioning for τ_n derived from the digital twin. The robot hand is parameterized as $\mathcal{R} = \{\Theta, \mathcal{L}\}$, where $\Theta \in \mathbb{R}^j$ denotes the articulated configuration of j joint degrees of freedom (DoF) and $\mathcal{L}$ defines the geometry via link meshes. Each grasp candidate $g \in \mathcal{G}$ is defined by $(T_{\mathcal{R}}, \Theta)$, where $T_{\mathcal{R}} \in SE(3)$ denotes the global hand pose.

4.2 Digital Twin Representation

ScrubGrasp constructs an inference-time digital twin comprising instrument meshes $\mathcal{T}$ and supporting surface Π_s, aligned to the robot's coordinate frame Ψ_r. This representation enables three conditioning inputs for grasp synthesis: bounded hand pose sampling (ϕ_{BS}), graspable surface identification (ϕ_{GS}), and scene collision validation (ϕ_{SC}).

4.2.1 Instrument Digital Twin

To create digital twins of the instruments in the scene, *ScrubGrasp* first employs Segment Anything Model V3 (SAM 3) [19] on the input RGB-D image $\mathcal{I}$ to extract instrument segmentation masks. The prompt

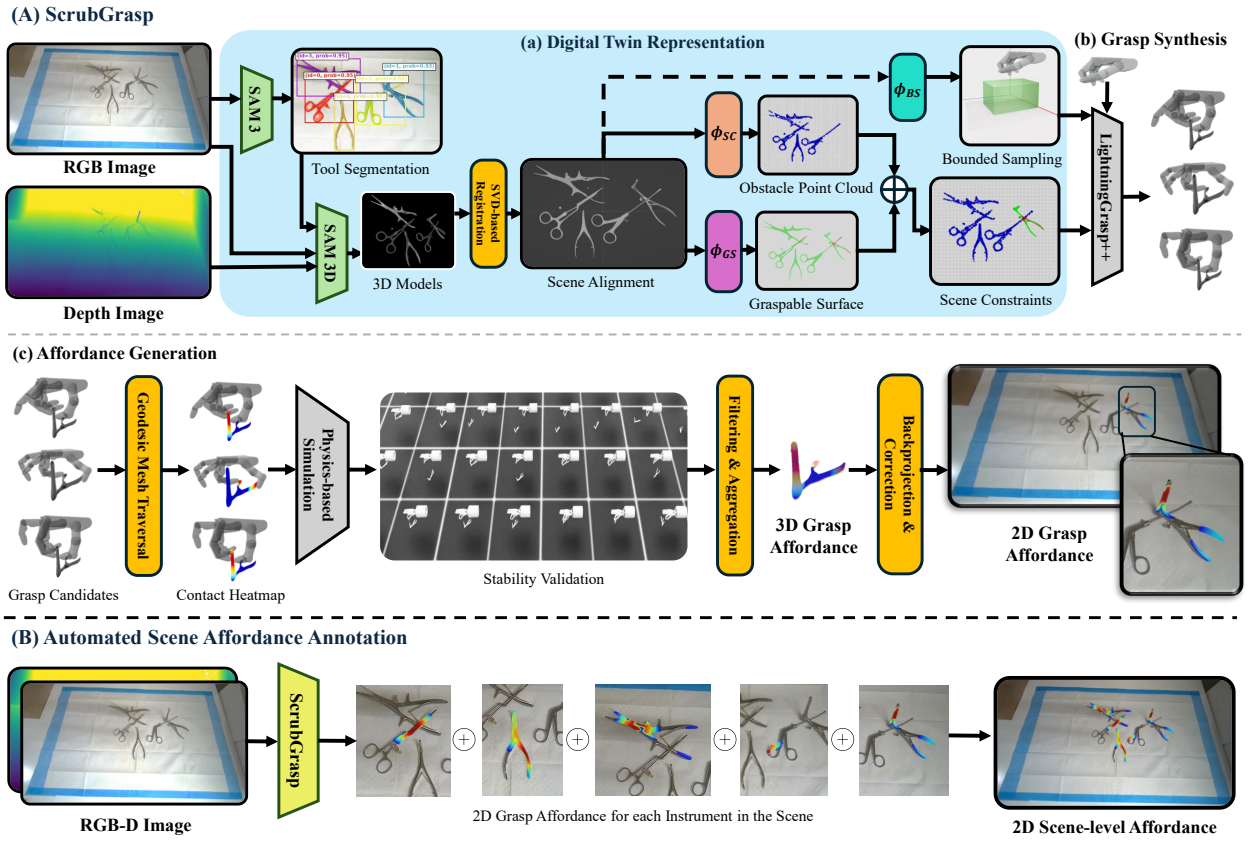


Figure 4.1: Overall methodology: (A) *ScrubGrasp*: (a) The pipeline takes an RGB-D image and uses SAM 3 and SAM 3D to generate the surgical tool meshes. It then performs scene alignment using SVD-based registration. (b) It conditions our backbone grasp generation model, modified Lightning Grasp, on bounded sampling (ϕ_{BS}), grasp surface conditioning (ϕ_{GS}), and scene collision conditioning (ϕ_{SC}) modules to produce grasp candidates. (c) These are then filtered via a physics engine to compute 3D and 2D grasp affordance using geodesic mesh traversal and backprojection, respectively. (B) This framework iteratively aggregates instrument affordances from RGB-D inputs into a scene-level annotation.

"tool" empirically yields robust masks across different scenes. These masks and $\mathcal{I}$ are then passed to SAM 3D Objects (SAM 3D) [42] to generate 3D mesh representations (digital twins) of the surgical instruments $\mathcal{T}$. The depth maps provided by $\mathcal{I}$ ensure high-fidelity reconstructions.

4.2.2 Scene Alignment

With SAM 3D predicting in the image coordinate frame, we align the instrument digital twin to the robot's coordinate frame Ψ_r, defined by a ground plane Π_r with normal $\mathbf{n}_r$. To recover the scene's orientation, we uniformly subsample vertices from all the instrument meshes $\mathcal{T}$ to form a sparse point cloud $\mathcal{P}$ and compute its centroid $\bar{\mathbf{p}}$. We then perform Singular Value Decomposition (SVD) [21] on the zero-centered matrix $\mathbf{A} = [\mathbf{p}_i - \bar{\mathbf{p}}]^\top \in \mathbb{R}^{|\mathcal{P}|\times 3}$. Since instruments lie on a planar surface, the direction of minimum

variance – the smallest singular value – yields the scene's ground normal $\mathbf{n}_s$. Global alignment is achieved by computing the rotation $\mathbf{R} \in SO(3)$ satisfying $\mathbf{R}\mathbf{n}_s = \mathbf{n}_r$. This registration approach removes the need for camera calibration, and the depth maps resolve scale ambiguity.

4.3 Grasp Synthesis

To enable grasp generation in multi-instrument environments and enforce physical and kinematic viability, we modify the Lightning Grasp backbone to integrate scene-level spatial conditioning.

4.3.1 Framework

Originally, Lightning Grasp [65] accepts an isolated object mesh $\mathcal{M}$ and a robot hand configuration $\mathcal{R}$ to produce a set of grasp candidates: $\mathcal{G} = \text{LightningGrasp}(\mathcal{M}, \mathcal{R})$. Operating in a robot-centric coordinate system Ψ_r for efficiency, it generates grasps by keeping $\mathcal{R}$ fixed at the origin while iteratively transforming $\mathcal{M}$ into various poses to find valid contact points. While this enables high-throughput inference, it creates a deployment gap: object-to-hand transformations generated in isolation can result in kinematically or physically implausible poses within a cluttered scene $\mathcal{I}$, ignoring spatial constraints from neighboring instruments $\mathcal{T}$ and the supporting surface Π_s.

To bridge this gap, we extend the framework into LightningGrasp++, which incorporates scene-level spatial awareness to ensure physical viability in multi-instrument environments. It takes three additional inputs: (a) a cuboidal bound within which a grasp can be sampled, (b) an occlusion mask, and (c) a scene point cloud through conditioning modules.

4.3.2 Conditioning Modules

Specifically, LightningGrasp++ generates grasps for a target instrument τ_n guided by a conditioning function $\phi(\tau_n)$. This function evaluates the scene-level constraints through three distinct modules:

$$\phi(\tau_n) = (\phi_{\text{BS}}(\tau_n, \mathcal{R}), \phi_{\text{GS}}(\tau_n \mid \mathcal{T}, \Pi_s), \phi_{\text{SC}}(\mathcal{R}, \mathcal{G} \mid \mathcal{T}, \Pi_s)) \tag{4.2}$$

where ϕ_{BS} (Bounded Sampling) constrains pose sampling to plausible approach directions, ϕ_{GS} (Graspable Surface Conditioning) ensures accessible grip areas by evaluating the spatial clearance of τ_n relative to neighboring instruments and the supporting surface, and ϕ_{SC} (Scene Collision Conditioning) enforces collision consistency by validating the articulated robot mesh against environmental obstacles.

Bounded Sampling (ϕ_{BS})

ϕ_{BS} re-parameterizes the generation space from the hand-centric canonical frame Ψ_r to the global scene coordinate system Ψ_s. Given a target instrument τ_n, we define its geometry relative to its geometric centroid, with a global pose $T_o \in SE(3)$ specifying its position and orientation in Ψ_s. To resolve discrepancies in axis orientation and handedness between the underlying model and the scene, we introduce a static change-of-basis transformation $T_{r\to s} \in SE(3)$, which aligns the canonical axes of Ψ_r with the world-centric axes of Ψ_s. Within this aligned subspace, we define a bounded cuboid region $\mathcal{B} \subset \Psi_r$ beneath $\mathcal{R}$, dynamically scaled to the τ_n dimensions and the $\mathcal{R}$'s link meshes $\mathcal{L}$. The backbone Lightning Grasp is constrained to

sample relative transformation T_g within $\mathcal{B}$ and successful grasp are projected back into the Ψ_s. The final global hand pose $T_{\mathcal{R}}$ is thus defined as:

$$T_{\mathcal{R}} = T_o \cdot T_{r \to s} \cdot T_g^{-1} \tag{4.3}$$

ensuring that locally identified grasps are mapped to the physical coordinates of τ_n in the world frame Ψ_s.

Graspable Surface Conditioning ($\phi_{\mathbf{GS}}$)

To enforce grasp synthesis on physically accessible instrument surface, we complement the instrument digital twin with scene-aware spatial constraints $\mathcal{I}$. We define a subset of graspable regions by evaluating the Euclidean distance between the target instrument τ_n and all other environmental entities. For each vertex v_i on the mesh of τ_n and any other object O_j in the set of obstacles $\mathcal{O} = \{\tau_j \in \mathcal{T} \mid j \neq n\} \cup \Pi_s$, we calculate the minimum distance:

$$d(v_i, O_j) = \min_{p \in O_j} \|v_i - p\| \tag{4.4}$$

A vertex v_i is considered graspable if $d(v_i, O_j) > \delta, \forall O_j \in \mathcal{O}$, where we empirically set $\delta = 0.5$ cm. This effectively prunes the input digital twin instrument mesh $\widehat{\tau_n} \subseteq \tau_n$ by removing regions occluded by, or in contact with, the ground plane or adjacent instruments. By conditioning the lightning grasp to generate grasp on filtered manifold $\mathcal{M} = \widehat{\tau_n}$, *ScrubGrasp* prevents grasp candidates in unreachable regions, improving success rate.

Scene Collision Conditioning ($\phi_{\mathbf{SC}}$)

Finally, we introduce scene collision conditioning to generate global scene-aware grasp candidates. We sample a dense point cloud $P_{\mathcal{O}}$ representing the environmental obstacles $\mathcal{O}$ from Ψ_s frame into Ψ_r frame through $T_{s \to r}$ transformation. A candidate $g \in \mathcal{G}$ is accepted as a feasible grasp only if the articulated robot links $\mathcal{L}$ do not penetrate the obstacle cloud, satisfying the condition:

$$\mathcal{L} \cap (T_g \cdot T_{s \to r} \cdot P_{\mathcal{O}}) = \emptyset \tag{4.5}$$

As an additional check against potential mesh projection errors, we enforce a geometric constraint to ensure that no individual robot link $l_k \in \mathcal{L}$ extends below the lowest vertical coordinate of the τ_n.

4.4 Affordance Generation

Translating discrete grasp candidates into learnable representations requires a comprehensive pipeline. Our approach sequentially constructs localized 3D interaction fields, aggregates them via rigorous dynamic stability verification, and finally projects these robust affordances into pixel-aligned, occlusion-aware 2D maps.

4.4.1 3D Grasp Affordance

Once physically viable candidate grasps have been generated, we can extract the contact points between the robot mesh and the object mesh through forward kinematics. To synthesize a dense contact heatmap on the target object corresponding to regions of valid robot-object interaction, we compute a continuous

geodesic confidence field over the object mesh. Initially, world-frame contact coordinates derived from the robot-object mesh intersections are transformed into the object's local coordinate space and projected onto the nearest respective vertices on the object mesh. The topological structure of the object is formulated as a sparse, undirected graph, where edge weights denote the Euclidean distances between connected vertices. We employ Dijkstra's algorithm to calculate the shortest-path geodesic distance, d, from the localized contact vertices to the surrounding mesh geometry. To ensure computational tractability, this search is strictly bounded by a maximum spatial threshold, d_{max}. The raw geodesic distances are subsequently mapped to an affordance score, $A(d) \in [0, 1]$, utilizing a normalized Gaussian decay formulation. This mapping ensures a peak affordance probability of exactly 1 at the direct point of robot contact and a smooth, monotonic decay to exactly 0 at the threshold boundary. The vertex-wise affordance score is formalized as:

$$A(d) = \frac{\exp\left(-\frac{d^2}{2\sigma^2}\right) - \exp\left(-\frac{d_{max}^2}{2\sigma^2}\right)}{1 - \exp\left(-\frac{d_{max}^2}{2\sigma^2}\right)} \tag{4.6}$$

where the floor term $\exp\left(-\frac{d_{max}^2}{2\sigma^2}\right)$ establishes the boundary condition at the cutoff distance, and σ parameterizes the spatial variance of the affordance decay. Vertices located beyond d_{max} are explicitly assigned an affordance score of zero. Finally, to account for multi-contact interactions, the global contact heatmap is synthesized by computing the supremum (vertex-wise maximum) of the affordance scores across all individual contact fields. This produces a unified, numerically stable contact representation across the object's surface topology.

To generate a meaningful 3D grasp affordance map, we compute and cumulate contact heatmaps from the grasp candidates produced by LightningGrasp++ for a specific tool τ_n in $\mathcal{I}$. However, LightningGrasp++ tests only for physical viability rather than dynamic stability. Thus, we first evaluate the candidates within a physics engine to filter out unstable grasps as detailed in Section 4.4.2. Following this verification, we aggregate the spatial influence across the entire distribution of stable robot-object interactions. For a given object, the localized geodesic confidence masks associated with each successful grasp are accumulated vertex-wise into a global spatial field. This cumulative field is then averaged to derive the mean expected affordance distribution over the mesh surface. To ensure that the final grasp representation is scale-invariant and strictly bounded within $[0, 1]$, we subsequently apply a global peak normalization. The averaged affordance map is scaled by its maximum vertex value, safeguarded against numerical instability by a marginal epsilon threshold. This yields a robust, standardized 3D probability density map that intrinsically highlights the most probabilistically favored grasp interaction regions across the object's geometry.

4.4.2 Grasp Candidate Stability Validation

While current grasp generation methods often lack physics-based validation for stability, *ScrubGrasp* addresses this limitation by leveraging a digital twin representation to rigorously evaluate candidates. Because initial generators like LightningGrasp++ assess only static physical viability (e.g., kinematics and collision avoidance), we integrate a mandatory simulation-based validation step using Isaac Lab (Section 5.4). For an initial set of K identified grasp candidates, dynamic stability is tested by applying external forces along six axis-aligned directions to the simulated object. Only the robust subset of k candidates ($k \subseteq K$) that successfully withstand these perturbations is retained. This physics-based filtering ensures that the subsequently aggregated 3D grasp affordance map is strictly grounded in interactions with the highest likelihood

of success in dynamic environments.

4.4.3 2D Grasp Affordance

To translate the aggregated 3D grasp affordance fields into the 2D visual domain, we systematically project the annotated object meshes onto the camera's image plane. Utilizing the camera intrinsic matrix K, the valid 3D mesh vertices are mapped to 2D pixel coordinates (u, v) via perspective projection. To robustly handle self-occlusions and inter-object occlusions within the scene, we employ a depth-aware visibility test where the projected vertices whose camera-space depth exceeds the spatially aligned observed depth map by a marginal tolerance ($\epsilon = 0.005$ m) are classified as occluded and subsequently discarded. As the direct projection of discrete vertices has minor reprojection errors, we apply a morphological dilation with an elliptical kernel to disseminate the valid affordance scores outward, followed by a Gaussian smoothing filter to resolve the dilated blocks into a continuous probability gradient. To ensure structural fidelity and prevent the affordance field from erroneously bleeding into the background or adjacent regions, this smoothed map is pointwise multiplied by the target object's binary segmentation mask, previously obtained via SAM 3. This generates 2D grasp affordance confidence mask A_{τ_n} for tool τ_n in $\mathcal{I}$.

4.5 Automated Scene Affordance Annotation

ScrubGrasp serves as an end-to-end framework that processes a raw scene $\mathcal{I}$ by first synthesizing a comprehensive digital twin, and subsequently generating physically grounded affordance representations for any given tool τ_n. By iterating this generation pipeline across the complete set of interactable tools $\mathcal{T} = \{\tau_1, \tau_2, \ldots, \tau_n\}$ present within the scene, the framework transitions from object-centric hypotheses to a holistic scene-level affordance mask. For each tool τ_n, the pipeline outputs an instance-level 2D affordance field $A_{\tau_n}(u, v) \in [0, 1]$, derived exclusively from dynamically stable grasps. To construct a unified representation for complex scenes containing multiple tools, these individual affordance fields are composited via a pixel-wise maximum operation:

$$A_{scene}(u, v) = \max_{\tau_n \in \mathcal{T}} A_{\tau_n}(u, v) \tag{4.7}$$

where (u, v) denotes the spatial pixel coordinates. This formulation ensures that overlapping interaction zones among multiple tools preserve their peak valid probability rather than artificially stacking. Ultimately, this engine produces a dense, globally consistent 2D grasp affordance map that is strictly grounded in physics simulation and perfectly aligned with the scene's original RGB-D observations.

Chapter 5

Experiments and Results

To rigorously assess the efficacy of *ScrubGrasp*, we design a comprehensive suite of experiments. Our evaluation is structured to address three core objectives. First, we benchmark our zero-shot grasp synthesis framework against state-of-the-art baselines on a custom-curated surgical instrument. Second, we validate the framework's practical utility by evaluating a downstream affordance prediction model bootstrapped entirely on annotations generated by *ScrubGrasp*, specifically assessing its ability to target unoccluded, context-aware graspable regions. Finally, we conduct extensive ablation studies to isolate the contributions of our core architectural conditioning modules. We also ablate our digital twin generation by evaluating SAM 3 and SAM 3D, assessing both 2D tool detection accuracy and 3D reconstruction fidelity against their respective computational scaling.

5.1 Experimental Setup

While our proposed pipeline is inherently agnostic to a specific robotic platform, we conduct our experiments using the Shadow Dexterous Hand [51]. This advanced anthropomorphic end-effector features 24 actuated joints, designed to mimic the kinematics and dexterity of the human hand. In our experiments, we impose kinematic constraints on these joints to simulate various lower-dimensional degree-of-freedom (DoF) configurations. When synthesizing grasps, we randomly sample from three distinct prehensile taxonomies: two-fingered (pinch), three-fingered (tripod), and fully actuated (palmar) grasps. For evaluation, we curate a custom dataset using ZED X camera to capture the scenes with ground-truth stereo images, depth maps, and intrinsic parameter, and utilized only the left image and its corresponding depth map for all evaluations. Furthermore, all high-fidelity, physics-based simulations are conducted within the Isaac Lab environment [34]. Both the simulation framework and the training of the downstream affordance estimation model are executed on an NVIDIA Quadro RTX 6000 GPU, which facilitates the highly efficient, parallelized evaluation of numerous grasp candidates.

5.2 Surgical Instrument Dataset

The clinical domain poses unique challenges for grasp synthesis: diverse instruments of varying shapes and sizes across surgical specialties, and varying levels of tool clutter throughout operative procedures.

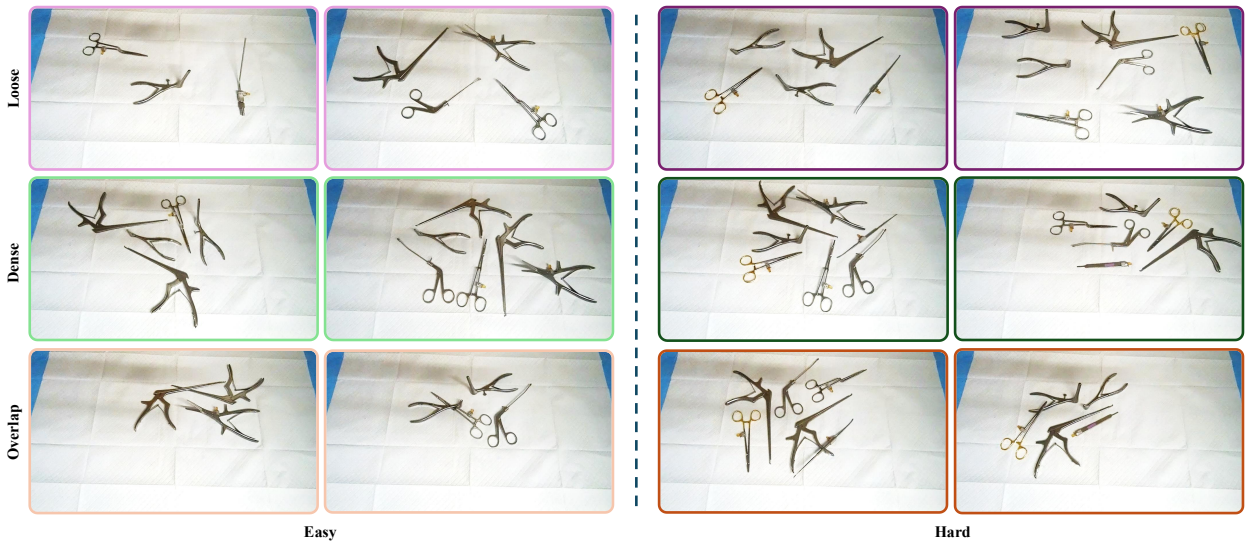


Figure 5.1: Sample scenes from our custom-curated dataset across six clutter configurations: loose (easy, hard), dense (easy, hard), and overlap (easy, hard). Easy scenes contain larger instruments with accessible, graspable regions, while hard scenes include thin, delicate instruments with complex geometries.

To systematically evaluate *ScrubGrasp* across these scenarios, we constructed a comprehensive evaluation dataset comprising six distinct configurations as shown in Figure 5.1. We defined three baseline levels of spatial clutter: loose, dense, and overlap. In a loose configuration, tools are scattered randomly across the surface. A dense configuration consists of tools placed in close proximity to one another. An overlap configuration features two to four physically overlapping tools within the scene. To account for challenges associated with variance in instrument complexity, we further subdivided each clutter category into easy and hard subsets. The easy subset consists of larger tools with accessible graspable regions, while the hard subset contains thin and delicate instruments. Easy scenes exclusively feature easily graspable tools, whereas hard scenes introduce a mixture of both easy and hard tools. Our final curated dataset comprises 20 scenes per subcategory, with each scene containing a random distribution of 3 to 8 tools drawn from a pool of 16 instruments (12 easy, 4 hard). In total, this yields 120 unique scenes encompassing over 600 distinct instrument poses for comprehensive grasp evaluation.

5.3 Evaluation Metrics

To quantitatively benchmark grasp synthesis methods, we evaluate the grasp success rate. A grasp candidate is deemed successful if it concurrently satisfies two criteria: (i) Collision-Free: We assess collisions by evaluating the intersection between the hand model and the scene's point cloud (derived from ground-truth depth maps using intrinsic camera parameters), ensuring no overlap with the supporting surface or neighboring tools, and (ii) Stability: Stability is evaluated by applying external force perturbations along all six axis-aligned directions ($\pm x, \pm y, \pm z$). A grasp is classified as stable only if the object displacement remains below 2 cm across all tested directions.

To evaluate the performance of our affordance prediction model, we employ three primary metrics:

Masked Mean Absolute Error (MAE), Top-1 Accuracy, and Top-5 Accuracy. To focus the evaluation on relevant regions rather than the background, we compute the MAE exclusively on areas where the ground truth affordance exceeds a minimum threshold ($\tau_{\text{mask}} = 0.05$). Let Y be the ground truth affordance map and $\hat{Y}$ be the predicted affordance map. The Masked MAE is defined as:

$$\text{Masked MAE} = \frac{1}{|M|} \sum_{i \in M} |\hat{y}_i - y_i| \tag{5.1}$$

where $M = \{i \mid y_i > 0.05\}$ represents the set of valid pixels for instrument grasping, and $|M|$ is the total number of pixels in this set.

Top-k accuracy evaluates the reliability of the model's highest-confidence predictions by cross-referencing their spatial locations against the ground-truth affordance map. Let N be the total number of evaluated images. For a given image n, let $\hat{Y}^{(n)}$ be the predicted affordance map and $Y^{(n)}$ be the corresponding ground truth. We define $\mathcal{P}_{n,k}$ as the set of pixel indices corresponding to the k highest predicted affordance scores in $\hat{Y}^{(n)}$. A prediction is considered a "hit" if at least one pixel in the predicted set $\mathcal{P}_{n,k}$ corresponds to a valid grasp point in the ground truth, defined as having an affordance value greater than or equal to a threshold $\tau_{\text{grasp}} = 0.5$. Using an indicator function $\mathbb{I}$, the Top-k accuracy is formally defined as:

$$\text{Top-}k \text{ Accuracy} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}\left(\max_{j \in \mathcal{P}_{n,k}} Y_j^{(n)} \geq \tau_{\text{grasp}} \right) \tag{5.2}$$

Specifically, we report two variations of this metric. Top-1 Accuracy ($k = 1$) evaluates the single most confident pixel prediction. Here, $\mathcal{P}_{n,1}$ contains only the index $i^* = \arg\max(\hat{Y}^{(n)})$, and the formulation simplifies to checking if the exact ground-truth value at that index satisfies $Y_{i^*}^{(n)} \geq 0.5$. Top-5 accuracy ($k = 5$) relaxes the spatial constraint by checking the top 5 highest-scored predicted pixels. A hit is recorded if any of these 5 indices point to a valid grasp region in the ground truth map.

To evaluate the digital twin representation, scene-level tool detection accuracy is reported, where a success is recorded only if all tools present in a given scene are correctly detected and identified. To measure the geometric accuracy of the reconstructed tools, we compute the Chamfer Distance (d_{CD}) between the normalized reprojected depth values (S_{pred}) and the normalized point clouds sampled from the ground-truth 3D models (S_{gt}). This is formulated as:

$$d_{CD}(S_{\text{pred}}, S_{\text{gt}}) = \frac{1}{|S_{\text{pred}}|} \sum_{x \in S_{\text{pred}}} \min_{y \in S_{\text{gt}}} \|x - y\|_2^2 + \frac{1}{|S_{\text{gt}}|} \sum_{y \in S_{\text{gt}}} \min_{x \in S_{\text{pred}}} \|x - y\|_2^2 \tag{5.3}$$

Finally, to assess the real-world applicability of the method, we report the inference time cost (TC). This represents the average latency, measured in seconds, across the various modular stages of our pipeline during inference.

5.4 Simulation-Based Grasp Validation

Given a robotic hand model, a target tool model, and a grasp candidate, we set up a simulation environment within Isaac Lab, a GPU-accelerated physics simulation platform built on NVIDIA Isaac Sim. To evaluate

Table 5.1: Grasp success rates (%) compared with state-of-the-art methods across six clutter configurations (loose, dense, overlap × easy, hard) and robotic hands with different joint degrees of freedom (DOF).

Method	Joint DOF	Loose		Dense		Overlap	
		Easy	Hard	Easy	Hard	Easy	Hard
DDGC [31]	7	22.03	13.04	29.45	22.47	17.09	17.87
DexGraspNet2 [69]	16	52.37	40.00	41.23	37.35	42.90	23.03
Lightning Grasp [65]	22	29.58	23.25	24.47	19.83	30.73	20.31
Ours	7	**55.17**	**34.33**	**59.67**	**40.24**	**38.56**	**38.04**
	16	**60.33**	**52.25**	**47.29**	**41.00**	**69.00**	**59.14**
	22	**57.00**	**53.71**	**47.25**	**50.0**	**68.00**	**50.29**

the success of a grasp candidate, the robotic hand is first positioned in the grasping pose around the target instrument. Surgical tools are initialized with a uniform density of 7.7 g/cm³ (stainless steel), enabling accurate computation of mass properties for complex geometries. Static and dynamic friction coefficients are set to 1.0, and gravity is disabled for both the manipulator and target tool to isolate kinematics and contact dynamics. We then assess the grasp stability by applying external force perturbations scaled to the tool's size along all six axis-aligned directions ($\pm$x, $\pm$y, $\pm$z). A grasp is deemed successful if the instruments displacement remains below 2 cm across all perturbation directions.

5.5 Comparison against State-of-the-Art Grasp Synthesis Methods

To the best of our knowledge, dexterous robotic grasping of surgical instruments remains unexplored in the literature. Therefore, we compare *ScrubGrasp* against state-of-the-art grasp synthesis methods designed for general objects: learning-based approaches (DDGC [31], DexGraspNet 2.0 [69]) and the optimization-based method (Lightning Grasp [65]) that *ScrubGrasp* extends upon. To determine whether models trained on generalized objects can successfully transfer to clinical setups, we perform inference on the baselines without any domain-specific fine-tuning. Since each baseline targets a distinct robotic hand– DDGC for a 7-DOF hand, DexGraspNet 2.0 for a 16-DOF hand, and Lightning Grasp for a 22-DOF hand– we ensure a fair comparison by adapting *ScrubGrasp* to match each hardware profile. Specifically, we constrain the joint limits of a simulated 24-DOF ShadowHand [51] to emulate the kinematics of the respective baseline end-effectors. The standard ShadowHand features 24 degrees of freedom (DOF): five in both the thumb and little finger, four in each of the remaining three fingers, and two in the wrist. To emulate the 22-DOF configuration for Lightning Grasp, we restrict only the wrist joints. To mimic the 16-DOF LEAP hand [45] for DexGraspNet 2.0, we completely immobilize the wrist and little finger, alongside one base knuckle joint of the thumb. Finally, to emulate the 7-DOF Barrett Hand [50] for DDGC, only the thumb, index, and middle fingers remain active. In this setup, the thumb is limited to two DOFs, while the index and middle fingers are configured to share the remaining five DOFs.

For quantitative evaluation, we employ weighted sampling on our custom dataset to extract 100 distinct instrument poses, proportionally distributed across six scene variations, and generate 10 grasps per target object – benchmarking each method on 1,000 synthesized grasps. Candidate selection varies based on the

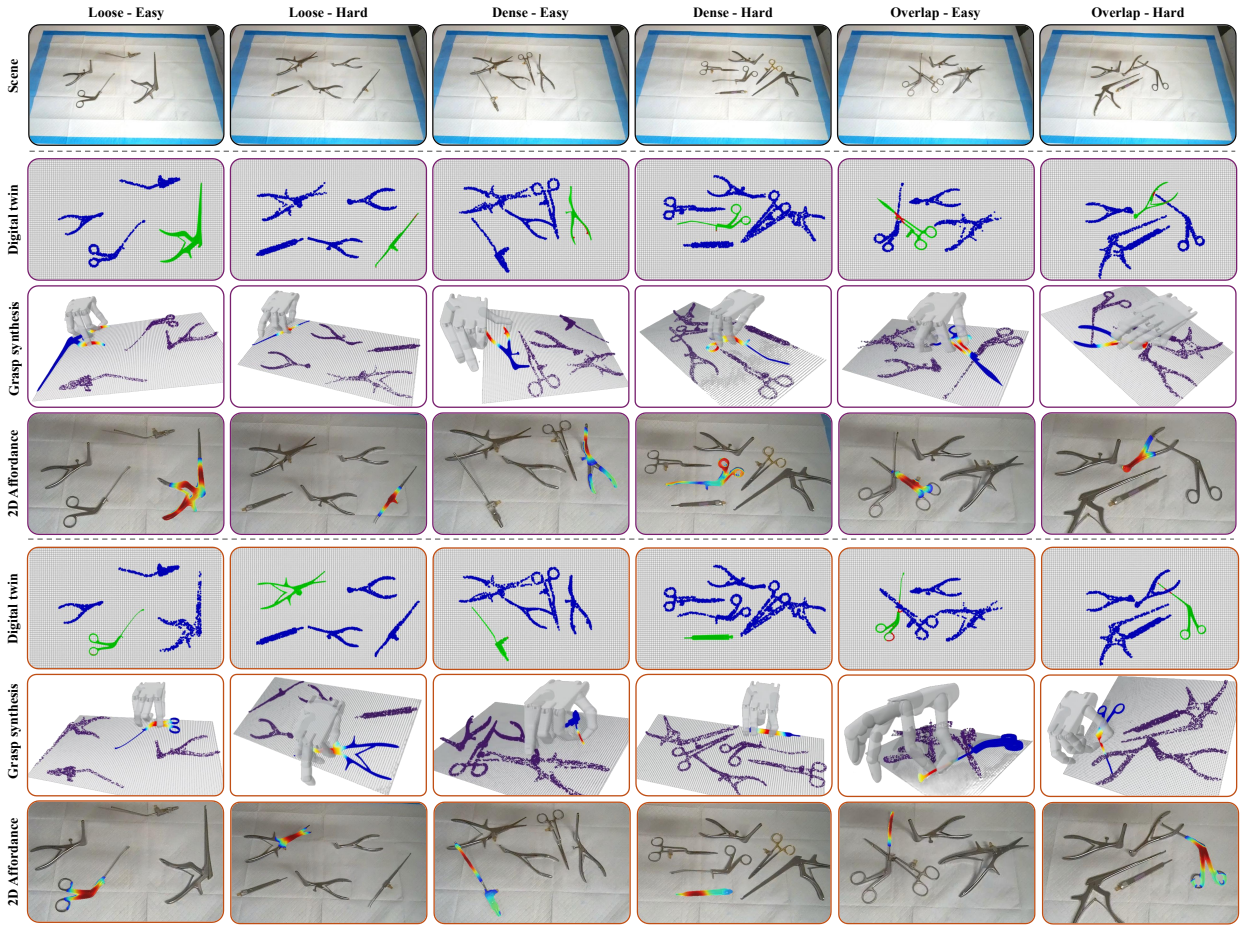


Figure 5.2: Qualitative results across six clutter configurations. For each target instrument: digital twin, synthesized grasp, contact heatmap, and 2D affordance. Digital twin: green (graspable regions), red (non-graspable due to occlusions, overlap, or surface contact), blue (neighboring obstacle point cloud). Contact and affordance heatmap: red (full graspability) to blue (≈ 0).

output structure of the respective methods. For DDGC [31] and DexGraspNet 2.0 [69], we select the 10 highest-ranked grasp candidates. In contrast, as LightningGrasp [65] does not provide a confidence ranking for its candidates, we randomly sample 10 grasps from the generated set for both LightningGrasp [65] and our *ScrubGrasp*.

As shown in Table 5.1, *ScrubGrasp* consistently outperforms all baselines across the six scene configurations and all hand DOF settings. We observe a significant generalization gap when deploying learning-based baselines (DDGC [31], DexGraspNet 2.0 [69]) on surgical instruments. Driven by a reliance on depth perception, these methods fail to accurately resolve the thin structural profiles of the tools, systematically inducing scene collisions. Lightning Grasp [65], while training-free, generates grasps without scene-level awareness, resulting in grasp candidates that potentially collide with the supporting surface or neighboring tools. In contrast, *ScrubGrasp*, with scene-context and graspable region conditioning, achieves robust

Table 5.2: Comparison of key method properties.

Method	Training free	Prior 3D Model free	Scene Conditioned	Simulation Enabled	Affordance Encoding
DDGC [31]	✗	✓	✓	✗	✗
DexGraspNet2 [69]	✗	✓	✓	✗	✓
LightningGrasp [65]	✓	✗	✗	✓	✗
Ours	✓	✓	✓	✓	✓

zero-shot performance across loose, dense, and overlapping configurations. Qualitative results in Figure 5.2 further illustrate how *ScrubGrasp* leverages digital twin representations to identify graspable regions and synthesize scene-aware grasps across diverse clutter configurations.

Table 5.2 summarizes key method properties. Both DDGC and DexGraspNet2 are prior 3D model-free and scene-conditioned, but they require training data and lack simulation capabilities. Between the two, only DexGraspNet2 incorporates affordance encoding through its training pipeline. Conversely, LightningGrasp is training-free and supports simulation, yet it relies on prior 3D models and lacks both scene conditioning and affordance encoding. In contrast, our *ScrubGrasp* uniquely satisfies all evaluated criteria: it enables training-free, prior 3D-model-free, and scene-conditioned grasp generation while successfully integrating grasp affordance generation. Furthermore, the inference-time digital twin representation makes our method simulation-enabled, potentially unlocking simulation-based validation of grasp candidates before real-world execution.

5.6 Annotation Validation via Downstream Affordance Prediction

While grasp synthesis can be directly employed for real-world applications, we further demonstrate its potential for curating synthetic data to train distilled vision models. To validate *ScrubGrasp* as an automated annotation generator, we trained a downstream affordance estimation network using the annotations generated by our framework on our custom-curated dataset. We model the affordance prediction task as a pixel-wise regression task problem using a U-Net [38] architecture. We utilize a ResNet-34 [14] backbone pre-trained on ImageNet [8] as the feature encoder. The network takes standard RGB images as input and outputs a single-channel affordance map, utilizing a Sigmoid activation function to constrain predictions to the $[0, 1]$ range. The dataset is divided into training (100 scenes) and validation (20 scenes) sets. Input images are resized and cropped to a resolution of 512×512 pixels and normalized using standard ImageNet statistics. During training, we apply several spatial data augmentations to prevent overfitting, including random shift, scale, and rotation (up to 30 degrees), as well as horizontal flips, vertical flips, and random resized cropping (scale 0.8 to 1.0). The model is optimized using the AdamW optimizer [29] with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We train the network for 150 epochs using a batch size of 8. We supervise the network using a custom Hybrid Affordance Loss, which combines Binary Cross-Entropy (BCE) and a spatially-weighted Smooth L1 Loss. The total loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{L1}}\mathcal{L}_{\text{SmoothL1}} \tag{5.4}$$

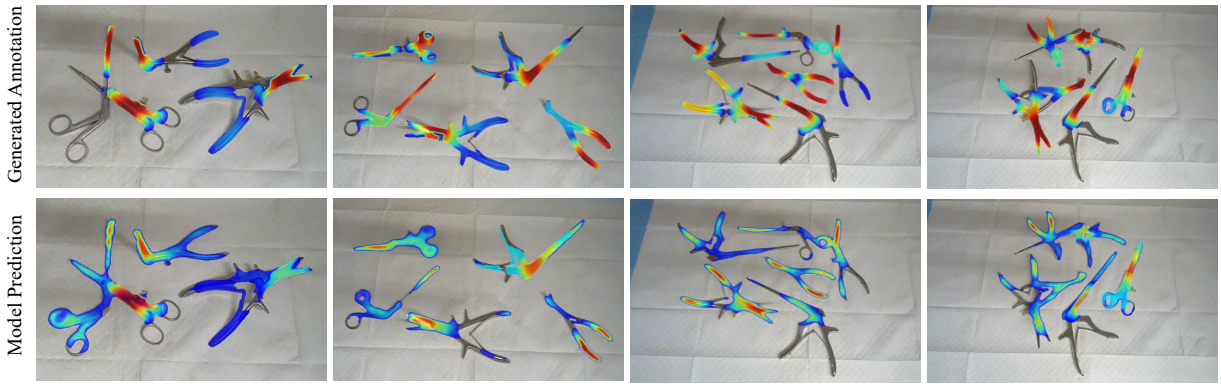


Figure 5.3: Qualitative comparison of ScrubGrasp affordance annotations vs. downstream model predictions on our custom dataset. Heatmap: red (full graspability) to blue (≈ 0).

where we empirically set $\lambda_{\text{BCE}} = 1.0$ and $\lambda_{\text{L1}} = 5.0$. To encourage the network to focus on highly affordant regions, the Smooth L1 loss component is multiplied by a spatial weight mask; specifically, a weight of 10.0 is applied to pixels where the target affordance is greater than 0.01, and a weight of 1.0 is applied elsewhere.

Qualitatively, the visual comparisons between the *ScrubGrasp*-generated labels and the network's predictions demonstrate a strong spatial correlation as shown in Figure 5.3. The model successfully isolates high-value grasp regions even in challenging, heavily occluded scenarios. Table 5.3 showcases the quantitative results. The network achieves a low Masked Mean Absolute Error (MAE) of 0.23, reflecting high-fidelity spatial continuity in its affordance estimation. Furthermore, the model yields a Top-1 Accuracy of 0.88 and a remarkable Top-5 Accuracy of 0.94. These strong performance metrics not only confirm that the chosen architecture can effectively resolve fine-grained grasp localizations, but more importantly, they substantiate the underlying *ScrubGrasp* pipeline. The exceptionally high Top-5 accuracy confirms that models trained on these automatically generated annotations can consistently identify multiple viable, redundant grasp candidates. These results substantiate *ScrubGrasp*'s utility as a highly reliable label generator for training downstream perception models to address computational complexity associated with an increasing number of instruments in the scene.

Table 5.3: Affordance prediction model performance on test set (20 scenes).

Metric	**Score**
Masked MAE	0.23
Top-1 Accuracy	0.88
Top-5 Accuracy	0.94

5.7 Ablation Studies

5.7.1 Ablation Study on Conditioning modules

Firstly, we conduct ablation studies to quantify the contribution of each conditioning module enabled by scalable digital twins in our *ScrubGrasp*. Building on Lightning Grasp (LG) [65], we incrementally add: (i) bounded sampling (BS), which constrains hand approach directions to the top cuboid region over the instrument; (ii) graspable surface conditioning (GS), which restricts contact points to accessible surface regions excluding regions that are occluded or in contact with supporting surfaces; and (iii) scene collision condi-

Table 5.4: Ablation study evaluating the contribution of each conditioning module. LG: Lightning Grasp [65] (base grasp generation), BS: bounded sampling, GS: graspable surface conditioning, SC: scene collision conditioning. SR: success rate (%), TC: time cost (seconds).

LG [65]	BS	GS	SC	SR ↑	TC ↓
✓				24.70	2.80
✓	✓			37.51	**1.95**
✓	✓	✓		39.14	2.27
✓	✓	✓	✓	**54.37**	2.66

tioning (SC), which conditions grasp synthesis to avoid collisions with neighboring tools and supporting surfaces in the scene. As shown in Table 5.4, each module contributes to improved grasp success rate (SR). The base Lightning Grasp achieves 24.70% SR, as it generates grasps without scene awareness. By eliminating grasps that approach from infeasible directions, BS improves SR to 37.51%. GS conditioning further increases SR to 39.14% by focusing on accessible regions on the instrument surface. Finally, integrating SC further increases the SR to 54.37% by rejecting candidates that collide with neighboring instruments or the ground plane. Integrating these modules also reduces time cost (TC). BS improves TC by 0.85s through early rejection of infeasible poses that would otherwise undergo full computation before being eliminated during post-filtering. While GS slightly increases TC, pruning ungraspable areas significantly reduces subsequent vertex-level computations. Although SC introduces localized overhead from point cloud sampling and filtering, the overall pipeline maintains computational efficiency while achieving a 2.2× increase in valid grasp generation.

5.7.2 Digital Twin Representation

We further ablate the performance of SAM 3 [19] and SAM 3D [42] employed within our *ScrubGrasp* in creating an inference time digital twin. We observe that SAM 3 demonstrates robust generalization to surgical scenes in our dataset, achieving an accuracy of 92.13% across all 120 scenes. It is also computation-friendly, with an average TC of 1.14 ± 0.003 seconds per image. In an accurate 3D digital twin representation of the instruments, SAM 3D achieved an average Chamfer distance of 0.011 across all tools detected in the dataset. We observe that the TC for SAM 3D scales linearly with the number of instruments in the scene (Figure 5.4), requiring an average of 16.17s to process a single tool.

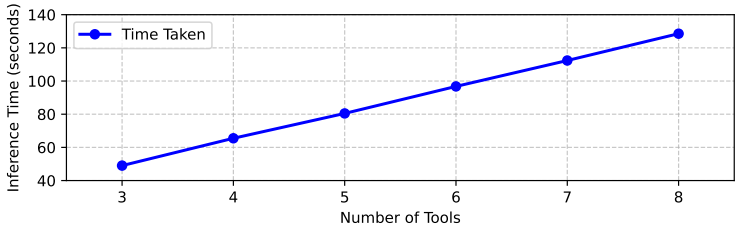


Figure 5.4: Ablation study on SAM 3 inference time in generating instrument digital twin based on total number of tools in the scene.

Chapter 6

Discussion

Building upon the proposed methodology and empirical results, we discuss potential downstream applications and acknowledge the current limitations of our framework. Furthermore, we outline promising avenues for future research.

6.1 Applications

The capabilities demonstrated by the *ScrubGrasp* framework– specifically its zero-shot adaptability and scalable data generation– open several highly impactful pathways for automation within and beyond the operating room.

The ability to accurately identify and grasp surgical instruments from cluttered, unstructured trays is the foundational requirement for robotic scrub nurses. By leveraging downstream policies trained via *ScrubGrasp*, these robotic assistants could autonomously locate, grasp, and hand off the correct instruments to the lead surgeon upon request, reducing cognitive load and shortening procedure times. After a procedure, surgical tools are typically deposited in disorganized, cluttered bins before being cleaned, sorted, and sterilized. *ScrubGrasp* can be deployed to automate the sorting process, seamlessly handling the wide morphological variety of instruments and chaotic spatial arrangements without requiring CAD models for every newly procured tool.

As *ScrubGrasp* constructs scene-aware digital twins from a single RGB-D image without requiring proprietary CAD models, it enables the rapid digitization of physical surgical environments. These high-fidelity virtual representations can be seamlessly integrated into Virtual Reality (VR) and Augmented Reality (AR) training simulators. Furthermore, they allow surgical teams to test different instrument layouts, robotic trajectories, and workflow optimizations in a risk-free virtual space before stepping into the physical operating room. In remote or tele-operated surgical setups, latency and limited depth perception can hinder a surgeon's ability to efficiently grasp tools. The 2D and 3D affordance heatmaps generated by *ScrubGrasp* can be integrated into AR displays or shared-control pipelines. This allows the robotic system to assist the remote surgeon by auto-aligning the end-effector with the optimal grasping region, ensuring a stable grip before manual manipulation resumes.

Beyond direct manipulation, *ScrubGrasp* functions as an invaluable data engine for medical machine learning research. Medical device manufacturers frequently introduce new tools, rendering existing datasets obsolete. By acting as an automated annotation pipeline, *ScrubGrasp* allows researchers and engineers to

rapidly bootstrap massive, physics-verified training datasets for new surgical tools in a matter of hours, entirely bypassing the prohibitive cost and time of manual human annotation.

6.2 Limitations

While this work represents a meaningful step toward scalable and robust automation of essential perioperative tasks, there are certain constraints to our current approach. Although our evaluation demonstrates robust grasping of complex and delicate instruments within a standardized simulation benchmark carefully calibrated to reflect real-world parameters, the framework has currently only been evaluated in a simulated environment. Given the strict safety tolerances of surgical settings, establishing a highly accurate simulated baseline was a necessary prerequisite. This allowed us to ensure foundational robustness and refine our control policies before introducing the dynamics and risks associated with a physical operating room.

Further constraints stem from our reliance on SAM 3 and SAM 3D for digital twin construction. The highly specular surfaces and complex, narrow geometries typical of surgical instruments can occasionally induce reconstruction failures, particularly in heavily cluttered scenes. We extensively explored alternative modeling techniques to mitigate this, including photogrammetry and traditional 3D scanning. However, these methods performed noticeably worse when handling the extreme metallic reflections and occlusions inherent to surgical trays. Consequently, SAM-based approaches remain our most viable and accurate option, despite these occasional artifacts. Additionally, the current inference architecture of SAM 3D causes computational costs to scale linearly with the number of instruments present, which can temporarily bottleneck execution speeds in highly dense scenes.

Finally, our current affordance annotations are intentionally restricted to scene-dependent graspable regions rather than localized semantic functional parts (e.g., the ‘handle’ or ‘tip’). We initially considered incorporating basic semantic labeling to facilitate instrument handovers. However, extensive consultations with clinicians revealed that perioperative instrument passing adheres to stringent, highly dynamic safety and ergonomic regulations. For instance, safely passing a scalpel requires grasping it such that the blade faces downwards and away from the receiver, while presenting the handle at a precise angle to be placed firmly into the surgeon’s palm without requiring repositioning. Simple part-based affordances proved insufficient to intuitively capture these complex, protocol-driven nuances, leading us to focus primarily on stable, scene-aware geometric grasping for this initial framework.

6.3 Future Work

Through this framework, a foundation is established at the intersection of robotic affordance generation and ambient intelligence in healthcare. As an immediate next step, *ScrubGrasp* needs to be validated on a physical robotic system. It is anticipated that the simulation calibration will facilitate sim-to-real transferability, enabling the practical real-world viability of both the proposed zero-shot grasp synthesis and the downstream policy distilled from the affordance annotations to be demonstrated.

To further improve the robustness of the digital twins, the perception pipeline can be refined to more effectively handle specular reflections and dense clutter. Technical efforts should also be directed toward optimizing the digital twin formulation so that real-time execution can be achieved and scaled efficiently, regardless of the volume of instruments on the tray.

Additionally, the scope of this research can be expanded into the domain of surgical human-robot interaction (HRI). By building upon the foundational affordance models, the complex, protocol-driven requirements of semantic instrument passing will be addressed to ensure that robotic handovers are performed safely and aligned ergonomically with clinical expectations. Ultimately, multimodal inputs– such as natural language instructions and semantics derived from diverse medical data formats– could be integrated in future iterations. This integration would allow more complex, context-aware actions to be executed by the system within the surgical environment.

Chapter 7

Conclusion

We presented *ScrubGrasp*, a zero-shot grasp synthesis framework that generates scene- and geometry-conditioned grasp candidates for surgical instruments from a single RGB-D image. By employing vision foundation models to generate digital twins, our methodology supports contextually adaptive grasp synthesis completely independent of offline training data, exact CAD specifications, and camera calibration. Extending beyond zero-shot execution, we demonstrated the framework's capacity to act as an automated scene-level annotation generator, producing 2D and 3D affordances by pairing simulation-verified stable grasps with graspable region heatmaps. We rigorously evaluated this pipeline on a newly introduced, custom-curated benchmark comprising 16 distinct surgical instruments across 120 varied scene orientations and over 600 unique poses. Under these challenging, cluttered conditions, *ScrubGrasp* demonstrated robust generalization across diverse tools and spatial arrangements, outperforming state-of-the-art grasp synthesis baselines by a significant margin of 2.2x. We further extended the custom-curated dataset by generating affordance annotations using our framework and demonstrated successful training of a downstream U-Net affordance policy. The resulting model effectively learned to target unoccluded graspable regions amidst clutter, proving that the generated affordances are both geometrically accurate and physically viable for downstream execution. This validates that *ScrubGrasp* serves as a scalable solution capable of bypassing labor-intensive data labeling, empowering the training of secure, end-to-end robotic models for safety-critical clinical environments.

Acknowledgements

I would like to thank Prof. Mathias Unberath for advising me during my time as a visiting graduate scholar at the ARCADE Lab at Johns Hopkins University, an experience that culminated in my master's thesis. I extend my gratitude to Dr. Lalithkumar Seenivasan, Han Zhang, and Dr. Roger D. Soberanis-Mukul, among many others, for their guidance throughout this project. I would also like to thank Prof. Siyu Tang for serving as my internal supervisor at ETH Zurich and taking the time to engage with this research. An extended thank you goes to Frano Rajič for his coordination during this process, and to Korrawe Karunratanakul for supervising my previous research projects at the VLG Lab.

This journey would not have been possible without all the pit stops and the people who supported me along the way. To my friends, my flatmates, and my peers in Zurich— particularly Sayan, Piyushi, Garima, Baidyanath, Justus, Leonie, Annika, and Saikiran— thank you for making ETH such a memorable place. Thank you to my Stryker peers, my Machine Vision and Sensing team, and my manager, Max Sirkin, for the opportunity to contribute to a worthwhile cause during my internship, and to Angi, her family, and Yugashree for making Freiburg, Germany, feel like home. Thank you to Dominik, Florence, Sristi, Nitik, my flatmates, and all the wonderful people I met (the list is too long) in Baltimore, USA during my time there.

A tree cannot grow if its roots are dry. To everyone back in India: 'thank you' is not enough for how you have molded my life, so I will simply acknowledge you here. To Ujjaini— the witness and alibi to my life for two decades and counting; to Viren, Yash, Prerna, Gauri, Diksha, Arunima, Chintan, Parthivi, Shuba, Achintya, Krithika, and to my partner Ansuman— the one who celebrates even my smallest wins. And finally, to my family— my mother, my father, my two grandmothers, Piku, Jamun, and Barfi— who raised me with a soul full of ambition and dreams and the determination to achieve them.

Appendix A

Supplementary Material

All ancillary data supporting this thesis– including the source code, datasets, experimental results, and multimedia demonstrations– are accessible at the following link: Supplementary Material

Bibliography

[1] Mohamed Mustaf Ahmed, Olalekan John Okesanya, Majd Oweidat, Zhinya Kawa Othman, Shuaibu Saidu Musa, and Don Eliseo Lucero-Prisno Iii. The ethics of data mining in healthcare: challenges, frameworks, and future directions. *BioData mining*, 18(1):47, 2025.

[2] Sonal Arora, Nick Sevdalis, Debra Nestel, Maria Woloshynowych, Ara Darzi, and Roger Kneebone. The impact of stress on surgical performance: a systematic review of the literature. *Surgery*, 147(3):318–330, 2010.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019.

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021.

[5] Zeyuan Chen, Qiyang Yan, Yuanpei Chen, Tianhao Wu, Jiyao Zhang, Zihan Ding, Jinzhou Li, Yaodong Yang, and Hao Dong. Clutterdexgrasp: A sim-to-real system for general dexterous grasping in cluttered scenes. *arXiv preprint arXiv:2506.14317*, 2025.

[6] Sue Min Cho, Jan Emily Mangulabnan, Han Zhang, Zhekai Mao, Yufan He, Pengfei Guo, Daguang Xu, Gregory Hager, Masaru Ishii, and Mathias Unberath. Humanoid robots as first assistants in endoscopic surgery, 2026.

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Hao Ding, Lalithkumar Seenivasan, Hongchao Shu, Grayson Byrd, Han Zhang, Pu Xiao, Juan Antonio Barragan, Russell H. Taylor, Peter Kazanzides, and Mathias Unberath. Towards Robust Automation of Surgical Systems via Digital Twin-based Scene Representations from Foundation Models, September 2024. arXiv:2409.13107 [cs].

[10] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, page 1–5. IEEE Press, 2018.

[11] Carlo Ferrari, John Canny, et al. Planning optimal grasps. In *Proceedings., 1992 IEEE International Conference on Robotics and Automation, 1992.*, volume 3, pages 2290–2295. IEEE, 1992.

[12] Association for the Advancement of Medical Instrumentation et al. *ANSI/AAMI ST79: 2017: Comprehensive Guide to Steam Sterilization and Sterility Assurance in Health Care Facilities*. AAMI, 2017.

[13] Atul A Gawande, David M Studdert, E John Orav, Troyen A Brennan, and Michael J Zinner. Risk factors for retained instruments and sponges after surgery. *New England Journal of Medicine*, 348(3):229–235, 2003.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Jonas Hein, Frédéric Giraud, Lilian Calvet, Alexander Schwarz, Nicola Alessandro Cavalcanti, Sergey Prokudin, Mazda Farshad, Siyu Tang, Marc Pollefeys, Fabio Carrillo, and Philipp Fürnstahl. Creating a Digital Twin of Spinal Surgery: A Proof of Concept. pages 2355–2364, 2024.

[16] Xuemin Hu, Shen Li, Tingyu Huang, Bo Tang, Rouxing Huai, and Long Chen. How Simulation Helps Autonomous Driving: A Survey of Sim2real, Digital Twins, and Parallel Intelligence. *IEEE Transactions on Intelligent Vehicles*, 9(1):593–612, January 2024.

[17] Janine Jagger, Jane Perry, Ahmed Gomaa, and Elayne Kornblatt Phillips. The impact of us policies to protect healthcare workers from bloodborne pathogens: the critical role of safety-engineered devices. *Journal of infection and public health*, 1(2):62–71, 2008.

[18] Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, and Jian Liu. Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14713–14724, 2023.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. pages 4015–4026, 2023.

[20] Constantin Kleinbeck, Han Zhang, Benjamin D. Killeen, Daniel Roth, and Mathias Unberath. Neural digital twins: reconstructing complex medical environments for spatial planning in virtual reality. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1301–1312, July 2024.

[21] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980.

[22] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024.

[23] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10581–10591, 2025.

[24] Junfei Li and Simon X. Yang. Digital twins to embodied artificial intelligence: review and perspective. *Intelligence & Robotics*, 5(1):202–227, March 2025. Publisher: OAE Publishing Inc.

[25] Puhao Li, Tengyu Liu, Yuyang Li, Yiran Geng, Yixin Zhu, Yaodong Yang, and Siyuan Huang. Gendexgrasp: Generalizable dexterous grasping. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 8068–8074. IEEE, 2023.

[26] Shunlei Li, Jin Wang, Rui Dai, Wanyu Ma, Wing Yin Ng, Yingbai Hu, and Zheng Li. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. *arXiv preprint arXiv:2409.19590*, 2024.

[27] Yiming Li, Wei Wei, Daheng Li, Peng Wang, Wanyi Li, and Jun Zhong. Hgc-net: Deep anthropomorphic hand grasping in clutter. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 714–720, 2022.

[28] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1):470–477, 2022.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Tyler Ga Wei Lum, Olivia Y. Lee, C. Karen Liu, and Jeannette Bohg. Crossing the Human-Robot Embodiment Gap with Sim-to-Real RL using One Human Demonstration, August 2025. arXiv:2504.12609 [cs].

[31] Jens Lundell, Francesco Verdoja, and Ville Kyrki. Ddgc: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters*, 6(4):6899–6906, 2021.

[32] Aiza Maksutova, Lalithkumar Seenivasan, Hao Ding, Jiru Xu, Chenhao Yu, Chenyan Jing, Yiqing Shen, and Mathias Unberath. Affordtissue: Dense affordance prediction for tool-action specific tissue interaction. *arXiv preprint arXiv:2604.01371*, 2026.

[33] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004.

[34] Mayank Mittal et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025.

[35] Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6813–6823, October 2021.

[36] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1374–1381. IEEE, 2015.

[37] Alejandra Perez, Han Zhang, Yu-Chun Ku, Lalithkumar Seenivasan, Roger Soberanis, Jose L. Porras, Richard Day, Jeff Jopling, Peter Najjar, and Mathias Unberath. Privacy-Preserving Operating Room Workflow Analysis using Digital Twins, July 2025. arXiv:2504.12552 [cs].

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[39] Peter Rubak, Ann-Eva Christensen, Mads Granlie, and Karin Bundgaard. Surgical instrument tray optimization process at a university hospital: A comprehensive overview. *Surgery Open Science*, 21:60–65, 2024.

[40] William Anthony Rutala and David Jay Weber. Guideline for disinfection and sterilization in healthcare facilities, 2008. 2008.

[41] Occupational Safety, Health Administration, et al. Occupational exposure to bloodborne pathogens; needlestick and other sharps injuries; final rule. *Fed Regist*, 66(12):5317–5325, 2001.

[42] SAM 3D Team et al. Sam 3d: 3dfy anything in images. 2025.

[43] Raghavasimhan Sankaranarayanan, Paul Stuart, Nicholas Ahn, Arno Sungarian, and Yash Chitalia. Towards autonomous instrument tray assembly for sterile processing applications. *arXiv preprint arXiv:2602.01679*, 2026.

[44] Rose Seavey. High-level disinfection, sterilization, and antisepsis: current issues in reprocessing medical and surgical instruments. *American journal of infection control*, 41(5):S111–S117, 2013.

[45] Kenneth Shaw, Ananye Agarwal, and Deepak Pathak. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. In *Robotics: Science and Systems (RSS)*, 2023.

[46] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

[47] Ritvik Singh, Arthur Allshire, Ankur Handa, Nathan Ratliff, and Karl Van Wyk. Dextrah-rgb: Visuomotor policies to grasp anything with dexterous hands. *arXiv preprint arXiv:2412.01791*, 2024.

[48] Stanislaw PA Stawicki, Susan D Moffatt-Bruce, Hesham M Ahmed, Harry L Anderson III, Tara M Balija, Irina Bernescu, Liza Chan, Laurie Chowayou, James Cipolla, Susette M Coyle, et al. Retained surgical items: a problem yet to be solved. *Journal of the American College of Surgeons*, 216(1):15–22, 2013.

[49] Yingbo Tang, Shuaike Zhang, Xiaoshuai Hao, Pengwei Wang, Jianlong Wu, Zhongyuan Wang, and Shanghang Zhang. Affordgrasp: In-context affordance reasoning for open-vocabulary task-oriented grasping in clutter. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9433–9439. IEEE, 2025.

[50] William T Townsend. The BarrettHand grasper–programmably flexible part handling and assembly. *Industrial Robot: An International Journal*, 27(3):181–188, 2000.

[51] Paul Tuffield and Hugo Elias. The shadow robot mimics human actions. *Industrial Robot: An International Journal*, 30:56–60, 02 2003.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[53] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[54] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366, 2023.

[55] Sizhe Wang, Yifan Yang, Yongkang Luo, Daheng Li, Wei Wei, Yan Zhang, Peiying Hu, Yunjin Fu, Haonan Duan, Jia Sun, et al. Scaleadfg: Affordance-based dexterous functional grasping via scalable dataset. *IEEE Robotics and Automation Letters*, 2025.

[56] Yi-Lin Wei, Zhexi Luo, Yuhao Lin, Mu Lin, Zhizhao Liang, Shuoyu Chen, and Wei-Shi Zheng. OmniDexGrasp: Generalizable Dexterous Grasping via Foundation Model and Force Feedback, October 2025. arXiv:2510.23119 [cs].

[57] Thomas G Weiser, Alex B Haynes, George Molina, Stuart R Lipsitz, Micaela M Esquivel, Tarsicio Uribe-Leitz, Rui Fu, Tej Azad, Tiffany E Chao, William R Berry, et al. Size and distribution of the global volume of surgery in 2012. *Bulletin of the World Health Organization*, 94(3):201, 2016.

[58] Cordula M Wetzel, Roger L Kneebone, Maria Woloshynowych, Debra Nestel, Krishna Moorthy, Jane Kidd, and Ara Darzi. The effects of stress on surgical performance. *The American Journal of Surgery*, 191(1):5–10, 2006.

[59] Turner Whitted. An improved illumination model for shaded display. In *ACM Siggraph 2005 Courses*, pages 4–es. 2005.

[60] Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, et al. Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11980–11990, 2025.

[61] Nan Xiao, Yunxin Fan, Farong Wang, and Fei Liu. arg-vu: Affordance reasoning with physics-aware 3d geometry for visual understanding in robotic surgery. *arXiv preprint arXiv:2603.26814*, 2026.

[62] Yi Xu, Ying Mao, Xianqiao Tong, Huan Tan, Weston B. Griffin, Balajee Kannan, and Lynn A. DeRose. Robotic handling of surgical instruments in a cluttered tray. *IEEE Transactions on Automation Science and Engineering*, 12(2):775–780, 2015.

[63] Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4737–4746, 2023.

[64] Jianglong Ye, Keyi Wang, Chengjing Yuan, Ruihan Yang, Yiquan Li, Jiyue Zhu, Yuzhe Qin, Xueyan Zou, and Xiaolong Wang. Dex1B: Learning with 1B Demonstrations for Dexterous Manipulation, June 2025. arXiv:2506.17198 [cs].

[65] Zhao-Heng Yin and Pieter Abbeel. Lightning grasp: High performance procedural grasp synthesis with contact fields. *arXiv preprint arXiv:2511.07418*, 2025.

[66] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, 41(7):690–705, 2022.

[67] Han Zhang, Lalithkumar Seenivasan, Jose L. Porras, Roger D. Soberanis-Mukul, Hao Ding, Hongchao Shu, Benjamin D. Killeen, Ankita Ghosh, Lonny Yarmus, Masaru Ishii, Angela Christine Argento, and Mathias Unberath. Did you just see that? Arbitrary view synthesis for egocentric replay of operating room workflows from ambient sensors, October 2025. arXiv:2510.04802 [cs].

[68] Han Zhang, Yiqing Shen, Roger D. Soberanis-Mukul, Ankita Ghosh, Hao Ding, Lalithkumar Seenivasan, Jose L. Porras, Zhekai Mao, Chenjia Li, Wenjie Xiao, Lonny Yarmus, Angela Christine Argento, Masaru Ishii, and Mathias Unberath. TwinOR: Photorealistic Digital Twins of Dynamic Operating Rooms for Embodied AI Research, November 2025. arXiv:2511.07412 [cs].

[69] Jialiang Zhang, Haoran Liu, Danshi Li, XinQiang Yu, Haoran Geng, Yufei Ding, Jiayi Chen, and He Wang. Dexgraspnet 2.0: Learning generative dexterous grasping in large-scale synthetic cluttered scenes. In *8th Annual Conference on Robot Learning*, 2024.

[70] Jiyao Zhang, Zhiyuan Ma, Tianhao Wu, Zeyuan Chen, and Hao Dong. Cadgrasp: Learning contact and collision aware general dexterous grasping in cluttered scenes. *arXiv preprint arXiv:2601.15039*, 2026.

[71] Lei Zhang, Kaixin Bai, Guowen Huang, Zhenshan Bing, Zhaopeng Chen, Alois Knoll, and Jianwei Zhang. Contactdexnet: Multi-fingered robotic hand grasping in cluttered environments through hand-object contact semantic mapping. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8495–8502. IEEE, 2025.

[72] Lei Zhang, Diwen Zheng, Kaixin Bai, Zhenshan Bing, Zoltan-Csaba Marton, Zhaopeng Chen, Alois Christian Knoll, and Jianwei Zhang. OmniDexVLG: Learning Dexterous Grasp Generation from Vision Language Model-Guided Grasp Semantics, Taxonomy and Functional Affordance, December 2025. arXiv:2512.03874 [cs].

[73] Yiming Zhong, Qi Jiang, Jingyi Yu, and Yuexin Ma. Dexgrasp anything: Towards universal robotic dexterous grasping with physics awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.