SMOKEBOT: A PERCEPTION-TO-MANIPULATION FRAMEWORK AND SIMULATION PLATFORM FOR ROBOTIC MANIPULATION IN SMOKE SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotic manipulation in smoke scenes is common yet highly challenging, as airborne scattering media degrade RGB observations and make active depth sensing unreliable. Existing robotic manipulation systems are primarily optimized for unobscured visual conditions, yet they overlook the complex multi-modal degradation induced by dynamic and spatially non-uniform smoke, which can severely impair perceptual reliability and thereby lead to a significant drop in manipulation performance.To address this challenge, we propose **SmokeManip**, a perception-to-manipulation framework for smoke-robust robotic manipulation without relying on active depth sensors. It decouples robotic perception into a scene understanding model and an affordance prediction model, both trained with geometric supervision to enable adaptive geometric and visual feature modeling under smoke. The former parses the task scene and localizes target objects, conditioning the latter to predict object-centric affordances that generate task-aware and physically feasible end-effector poses. For manipulation, SmokeManip further leverages a reusable action primitive library and hierarchical closed-loop recovery mechanism to compensate for perception inaccuracies and maintain stable manipulation. Beyond the algorithmic challenges, smoke-scene data collection and robotic manipulation evaluation lack a suitable platform. Therefore, we further construct **SmokeLab**, a dedicated simulation platform for robotic manipulation in smoke scenes. SmokeLab provides high-fidelity dynamic smoke simulation, automated data collection, and unified evaluation protocols for robotic perception and manipulation. Extensive experiments in SmokeLab and real-world demonstrate that our approach outperforms existing baselines across varying smoke densities. The project page is available at `https://smokebot-web.github.io/`.

1 INTRODUCTION

As shown in Figure 1, smoke[1] is common in robotic manipulation scenarios: cooking fume rises in kitchens, water mist fills bathrooms, and dense smoke appears at fire scenes. Airborne scattering particles attenuate and redirect light along the imaging path, degrading RGB observations and making active depth sensing unreliable (Nguyen et al., 2017; Fujimura et al., 2020a). In particular, scattering media disrupt structured-light and time-of-flight depth sensors commonly used by robots, making their depth estimates unreliable (Fujimura et al., 2020b; Dhrafani et al., 2025). Existing robotic manipulation systems typically assume clean RGB and trustworthy depth, thereby overlooking the complex multimodal degradation induced by dynamic and spatially non-uniform smoke, which can severely impair perceptual reliability and lead to a significant drop in manipulation performance.

To address this problem, we propose **SmokeManip**, a perception-to-manipulation framework for smoke-degraded scenes that operates without relying on active depth sensors at deployment. At the perception stage, two specialized modules are used: **SmokeScene** localizes task-relevant objects and refines their masks from global and local RGB observations, while **SmokeAfford** predicts object-centric interaction regions from the local observation and target-object mask. Both modules

[1]Smoke, fog, fume, steam, mist, and dust are airborne scattering media differing in particle properties and formation mechanism. This paper does not distinguish these similar phenomena and uses *smoke* for simplicity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

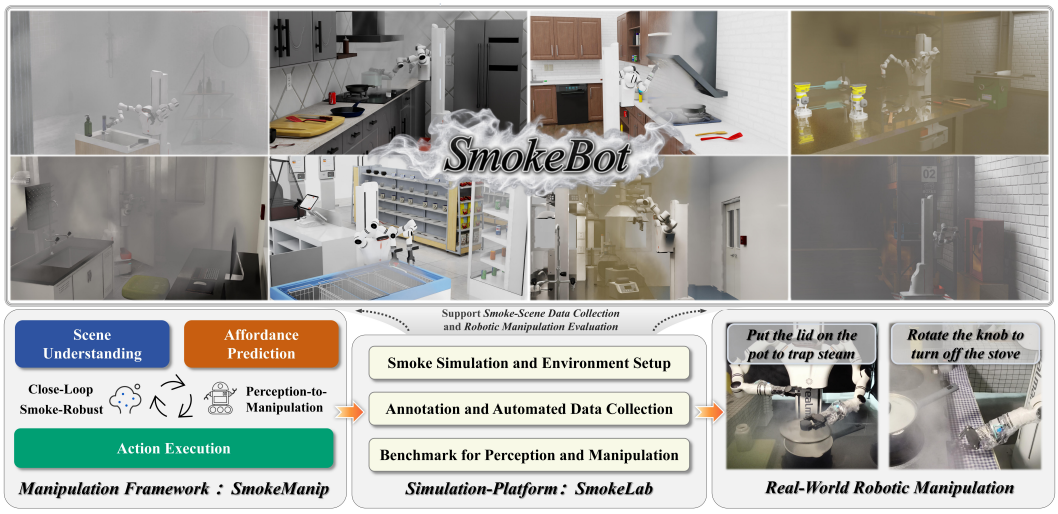


Figure 1: **Overview of SmokeBot**, comprising SmokeManip, a perception-to-manipulation framework, and SmokeLab, a simulation platform, both designed for robotic manipulation in smoke scenes. SmokeManip uses SmokeLab's smoke-centric scenes and tasks for data collection and evaluation before transfer to real-world smoke scenes for accurate perception and stable manipulation.

use geometric supervision to enable adaptive geometric and visual feature modeling under smoke. The resulting perception outputs are connected to action execution through hierarchical closed-loop recovery, which supports targeted re-observation and pose retry.

Beyond the algorithmic challenges, smoke-scene data collection and robotic manipulation evaluation lack a suitable platform. Atmospheric scattering formulas provide controlled image degradation and have been used for synthetic fog-scene modeling (Sakaridis et al., 2018; Dai et al., 2020); however, they can produce unrealistic smoke boundaries, weak volumetric appearance, and no support for dynamic robot execution. Therefore, we construct **SmokeLab**, a simulation benchmark for robotic perception and manipulation in smoke. SmokeLab provides high-fidelity dynamic smoke simulation under configurable spatial and temporal conditions, automated data collection, execution-oriented annotations, and unified protocols for perception, action primitives, and compositional manipulation.

We evaluate SmokeScene and SmokeAfford across smoke densities and SmokeManip on action primitives and compositional manipulation tasks in SmokeLab and real-world smoke scenes. Our approach outperforms existing baselines across varying smoke densities, while component ablations verify the decoupled design. In summary, our contributions include:

- We propose **SmokeManip**, a unified perception-to-manipulation framework that decouples smoke-robust perception from primitive-based execution and combines affordance-guided pose generation with hierarchical recovery for depth-free deployment.
- We develop **SmokeScene** and **SmokeAfford**, complementary perception models that localize task-relevant objects and predict object-conditioned interaction regions using geometry-guided features under spatially varying smoke.
- We introduce **SmokeLab**, a simulation platform and hierarchical benchmark with dynamic volumetric smoke, automated data collection, execution-oriented annotations, standardized primitive and compositional tasks, and unified evaluation protocols.
- We conduct comprehensive simulation and real-world evaluations, demonstrating consistent gains over perception and manipulation baselines and validating the contributions of perception, pose generation, and closed-loop execution.

2 Related Work

Robotic manipulation under degraded visibility. Some studies optimize robotic perception in low-light conditions via cross-modal knowledge distillation (Zhong et al., 2025; Zhu et al.,

2026a), visual enhancement (Niu et al., 2023; Zhang et al., 2023), and adaptive structural compensation (Zhong et al., 2026), to achieve more stable grasping execution. Recent studies focus on improving the performance of vision-language-action models under degraded visual observations. StableVLA (Fu et al., 2026) and CRT-VLA (Guarnizo Orjuela et al., 2026) introduce tailored adaptation modules to increase robustness to photometric degradation and visual corruption, while E-VLA (Zhai et al., 2026) and EventVLA (Yang et al., 2026) integrate event streams to complement visual cues in low-light and motion-blurred visual inputs. However, robotic manipulation under smoke remains underexplored. Smoke is dynamic and spatially nonuniform, and particle scattering interferes with both RGB and depth sensing, creating more complex and time-varying visual degradation. Therefore, we introduce SmokeManip, which combines smoke-robust visual perception with closed-loop execution for accurate and stable manipulation under smoke.

Robust perception in smoke scenes. Perception in smoke scenes has been extensively studied as a challenging task in computer vision (Dai et al., 2020; Liao et al., 2022; Liu et al., 2024; 2025a; Feng et al., 2026). Some works improve perception robustness by learning smoke-invariant features (Lee et al., 2022) or exploiting physical priors (Liu et al., 2025b). Other works further explore the potential of domain adaptation for robust visual perception (Ma et al., 2022; Li et al., 2023a; Jang et al., 2024; Wang et al., 2026). However, most of these works target autonomous driving, unmanned aerial vehicles, or remote sensing views. In contrast, perception models for robotic manipulation scenarios require a robot-centric view and must account for manipulation constraints and task semantics (Jia et al., 2025; Liu et al., 2026). In this paper, we develop robot-centric perception models tailored to robotic manipulation that achieve robust scene understanding and manipulation affordance prediction through geometric supervision.

Simulation benchmarks for robotic manipulation. Advances in simulation technologies have enabled controllable, reproducible, and efficient environments for robotic manipulation (Todorov et al., 2012; Coumans and Bai, 2021; Xiang et al., 2020; NVIDIA, 2026). Realistic physical simulation supports high-quality and large-scale data collection, with GPU-accelerated computation further improving scalability (James et al., 2020; Li et al., 2023b; Nasiriany et al., 2024; 2026). Some works provide platforms to enable fair evaluation across robotic manipulation tasks (Yu et al., 2020; Mees et al., 2022; Liu et al., 2023; Gu et al., 2023; Mu et al., 2025). Recent studies leverage simulation platforms to collect data for manipulating challenging objects (Xian et al., 2023; Wang et al., 2024; 2025; Cao et al., 2026; Zhu et al., 2026b) and performing specialized manipulation tasks (Li et al., 2025; Lan et al., 2026), alleviating the difficulty of real-world data acquisition. Despite these advances, a unified simulation platform dedicated to high-fidelity dynamic smoke and smoke-specific manipulation tasks is still lacking. To fill this gap, we introduce SmokeLab, a unified platform for smoke-related perception and manipulation tasks with controllable volumetric smoke.

3 SMOKEMANIP

3.1 OVERVIEW OF SMOKEMANIP

Figure 2 gives an overview of the SmokeManip pipeline, from global-to-local perception through affordance-guided pose generation and primitive execution with recovery. SmokeManip combines SmokeScene, SmokeAfford, a pose generator, an action primitive library, and a closed-loop execution policy. SmokeScene uses a global RGB observation to localize the task-relevant target and a local observation to refine its target-object mask. Conditioned on the local observation, mask, and task specification, SmokeAfford predicts an object-centric affordance map.

The smoke-adapted depth estimator reconstructs local object geometry and lifts high-response affordance pixels into 3D. The pose generator combines geometric cues with affordance responses to produce physically feasible 6-DoF candidates. If the selected pose fails, the local observation is reused to recompute the affordance map and pose candidates. If SmokeScene cannot find a target mask, the system first performs global re-observation and restarts localization.

3.2 AFFORDANCE-GUIDED POSE GENERATION

SmokeManip first selects high-response 2D pixel positions from the affordance map. The pose generator leverages a smoke-adapted DepthAnythingv2 estimator fine-tuned on SmokeLab RGB–depth pairs around these pixels to produce 6-DoF contact and grasp candidate poses. Contact candidates

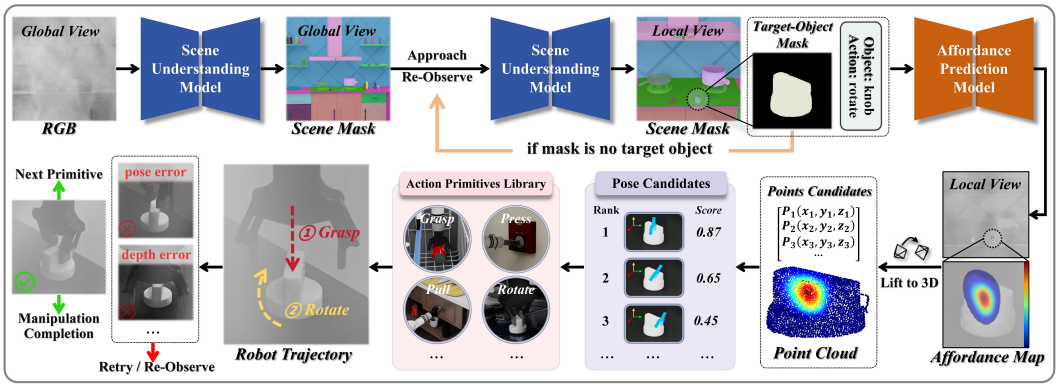


Figure 2: **Overview of SmokeManip**. SmokeScene parses the task scene and provides a target-object mask to SmokeAfford. SmokeAfford predicts object-centric affordances from local observations, after which pose candidates are ranked, executed through reusable action primitives, and verified with hierarchical closed-loop recovery during smoke manipulation.

are defined by reconstructed 3D points and their surface normals; grasp candidates follow the representation formulation of AnyGrasp (Fang et al., 2023). We rank contact and grasp candidates by combining their affordance heatmap responses with pose-generator confidence. The scoring function is defined in Appendix A.

3.3 Primitive-Based Action Execution

The action category indexes a ProMP-based (Paraschos et al., 2013) action primitive library. Each library entry stores a reusable Cartesian motion pattern together with its gripper schedule, action feasibility constraints, and object-state success predicate. Given the current end-effector starting pose and the selected grasp candidate pose, the primitive generates the corresponding Cartesian trajectory. The execution wrapper enforces gripper schedules and action feasibility constraints, and verifies the resulting object-state transition. More details are provided in Appendix A.2.

To address execution failures caused by perceptual inaccuracies in smoke scenarios, the framework adopts a hierarchical closed-loop recovery strategy. Under this mechanism, if a grasp candidate fails while the target remains successfully localized, the system either executes the next feasible grasp pose or acquires new local RGB observations to update perceptual outputs under dynamically varying smoke. When local observations are insufficient for reliable target identification, the robot conducts global re-observation and restarts object localization. Validated object-state transitions trigger the progression of subsequent task conditions.

4 SmokeScene and SmokeAfford

As shown in Fig. 2 and Fig. 3, SmokeManip decouples smoke-robust perception into two complementary, independently optimized models. SmokeScene processes global and local RGB observations to identify the target object and predict its mask and semantic category. SmokeAfford then combines the local RGB observation, target-object mask, and action category to predict an object-centric affordance map. Sections 4.1 and 4.2 describe the training objectives and feature transformations of the two perception models; neither requires active depth at deployment. The corresponding benchmark data flow and evaluation protocol are introduced in Section 5.3.

4.1 SmokeScene: Task-Oriented Scene Understanding

Overview of SmokeScene. Given a global or local RGB observation, SmokeScene produces object instance masks together with their corresponding semantic predictions. Although smoke degrades local appearance and object boundaries, geometric cues retain valid spatial-structural information under smoke (Feng et al., 2026). Accordingly, we employ DepthAnythingv2 as a frozen geometry teacher to provide geometric reference features.

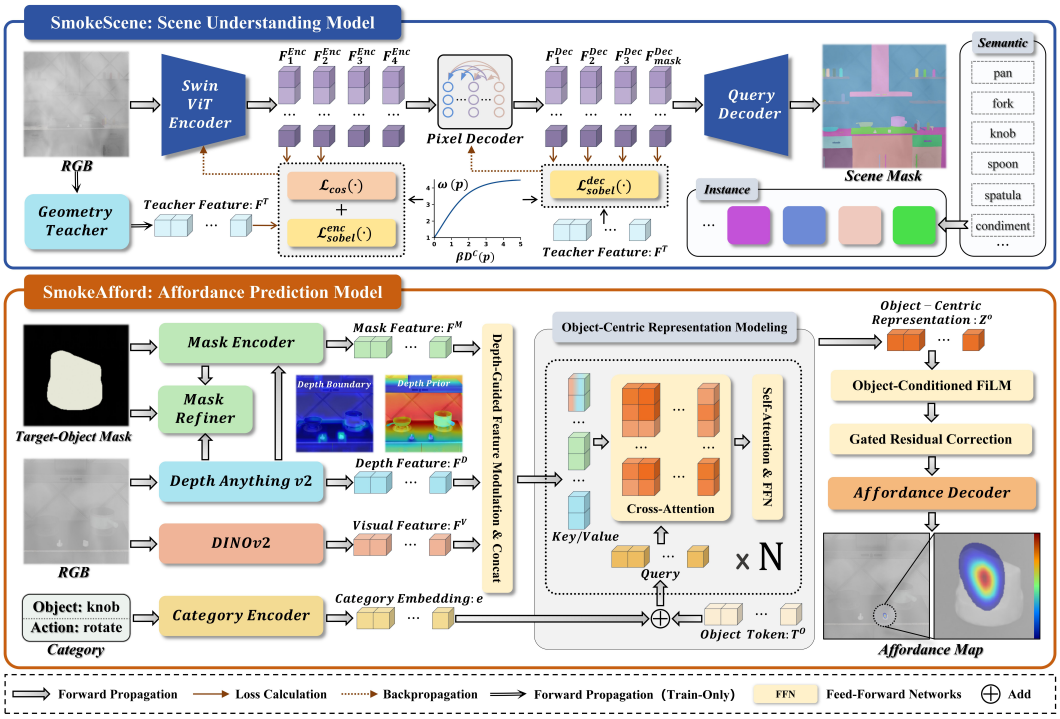


Figure 3: **Perception models used by SmokeManip**. SmokeScene performs task-oriented scene understanding from global and local RGB observations. SmokeAfford predicts an affordance map from the target object's local RGB observation. The two models are optimized separately, with the target-object mask passed explicitly during inference from SmokeScene to SmokeAfford.

Local smoke severity weight. Smoke is spatially non-uniform, so uniform pixel weighting can dilute supervision from heavily occluded regions. We derive a local severity signal $1-\exp(-\beta D^{\mathrm{C}}(p))$ from smoke density and clean depth, then use its normalized form to weight privileged geometry supervision. The resulting encoder and decoder weights, ω_ℓ and ω_k, emphasize smoke-affected regions; the full formulation is given in Appendix B.

Geometry supervision on encoder and decoder. Early high-resolution encoder features contain smoke-degraded visual responses and intrinsic geometric cues. Aligning intermediate features against geometry-teacher responses purifies geometry cues and suppresses texture-driven shortcut learning. We impose cosine and Sobel-gradient supervision on the first two encoder features. Since deeper encoder layers mainly yield semantic abstractions and lack fine pixel-level geometric cues, we additionally apply multi-scale Sobel-gradient supervision to the pixel-decoder features and mask feature. The complete encoder and decoder loss formulations are provided in Appendix B.

The total objective. The semantic–instance losses and the two geometry losses are combined to form the SmokeScene training objective:

$$\mathcal{L}_{\text{scene}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{ins}} + \alpha_{\text{geo}}(t)\left(\lambda_{\text{enc}}\mathcal{L}_{\text{enc-geo}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec-geo}}\right). \tag{1}$$

Here $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{ins}}$ are the semantic and instance losses; the full decomposition is given in Appendix B. The scalar $\alpha_{\text{geo}}(t)$ applies warm-up and late decay to the complete geometry objective, while the severity weight maps distribute that supervision spatially within each feature level.

4.2 SmokeAfford: Object-Centric Affordance Prediction

Overview of SmokeAfford. SmokeAfford predicts an action-specific affordance map. To guide manipulation toward the target object in cluttered operation scenes, it uses a target-object mask. Given a smoke RGB observation, the mask, and object and action categories, it extracts aligned visual, mask, and depth features F^{V}, F^{M}, and F^{D}. A category encoder combines the matched object and action categories with learned category embeddings to produce $\mathbf{e}$.

Object mask refinement and encoding. Smoke can make the input mask incomplete or shifted. A frozen DepthAnythingv2 module provides F^{D}, a depth prior D^{P}, and its boundary map E^{D}. The mask encoder encodes these cues with $\widehat{\mathcal{M}}$ into F^{M}, then refines the soft mask. To prevent smoke-induced mask errors from propagating to affordance prediction, we refine the target-object mask with mask geometry and depth-boundary cues before using it as a spatial prior. The residual refinement formulation and variable definitions are provided in Appendix A.

Depth-guided feature modulation. To compensate for smoke-degraded visual cues, we introduce a gated residual modulation to inject depth information while preserving the visual representation:

$$F^{\mathrm{P}} = \mathcal{F}_{\mathrm{P}}\big(\mathrm{Concat}\left(F^{\mathrm{V}}, F^{\mathrm{M}}\right)\big), \qquad \widetilde{F}^{\mathrm{P}} = F^{\mathrm{P}} + \alpha_{\mathrm{D}} G_{\mathrm{D}}\big(F^{\mathrm{P}}, F^{\mathrm{D}}, F^{\mathrm{M}}\big) \odot \Delta_{\mathrm{D}}\big(F^{\mathrm{P}}, F^{\mathrm{D}}\big). \quad (2)$$

Here, G_{D} predicts an element-wise gate, Δ_{D} produces the depth-guided residual, α_{D} controls its magnitude, and $\odot$ denotes element-wise multiplication.

Object-centric representation modeling. To model object-centric interactions conditioned on object and action semantics, we introduce learnable object tokens T^{O}. Conditioned by $\mathbf{e}$, these tokens attend to the fused visual, mask, and depth feature representations to produce an object-centric representation $\mathbf{Z}^{\mathrm{O}}$. To further couple object and action semantics with local visual-geometric cues, the conditioned tokens attend to the fused spatial features and form the representation:

$$\mathbf{Z}^{\mathrm{O}} = \mathcal{T}_{\mathrm{O}}\big(\text{Query} = T^{\mathrm{O}} + \mathbf{e},\ \text{Key=Value} = \mathrm{Concat}\left(\widetilde{F}^{\mathrm{P}}, F^{\mathrm{M}}, F^{\mathrm{D}}\right)\big). \quad (3)$$

Object-conditioned affordance decoding. To map the object and action semantics encoded in $\mathbf{Z}^{\mathrm{O}}$ to spatially precise action affordances, the decoder conditions the fused spatial features on this representation. It generates FiLM parameters and a gated residual correction to adapt $\widetilde{F}^{\mathrm{P}}$ to F^{A}:

$$F_0^{\mathrm{A}} = \widetilde{F}^{\mathrm{P}} \odot (1+\gamma) + \beta, \qquad F^{\mathrm{A}} = F_0^{\mathrm{A}} + \alpha_{\mathrm{P}} G_{\mathrm{P}}(F_0^{\mathrm{A}}, \mathbf{Z}^{\mathrm{O}}) \odot \Delta_{\mathrm{P}}(F_0^{\mathrm{A}}, \mathbf{Z}^{\mathrm{O}}). \quad (4)$$

Here, γ and β are FiLM scale and bias parameters generated from $\mathbf{Z}^{\mathrm{O}}$. The decoder then combines this feature, the refined mask, and the object representation to produce a spatially precise affordance map. The explicit output mapping and decoder definition are provided in Appendix A.

The total objective. To obtain accurate, distributions-aware affordance heatmaps, we formulate a composite learning objective for affordance prediction. The total affordance objective is

$$\mathcal{L}_{\mathrm{afford}} = \lambda_{\mathrm{bce}}\mathcal{L}_{\mathrm{bce}} + \lambda_{\mathrm{kld}}\mathcal{L}_{\mathrm{kld}} + \lambda_{\mathrm{edge}}\mathcal{L}_{\mathrm{edge}} + \mathcal{L}_{\mathrm{aux}}. \quad (5)$$

Specifically, $\mathcal{L}_{\mathrm{bce}}$ provides dense pixel-wise heatmap supervision, $\mathcal{L}_{\mathrm{kld}}$ aligns predicted and target affordance distributions, and $\mathcal{L}_{\mathrm{edge}}$ regularizes affordance boundaries. $\mathcal{L}_{\mathrm{aux}}$ adds consistency regularization for original and perturbed masks. The definitions of all loss terms are given in Appendix B.

5 SmokeLab

5.1 Smoke Simulation and Environment Setup

Smoke Simulation. As shown in Figure 4, SmokeLab adopts volumetric flow-based smoke simulation to reproduce authentic and dynamic smoke interference in real-world scenarios. Built upon NVIDIA Omniverse Flow, the platform formulates smoke as a sparse voxel field within the Isaac Sim engine. Unlike depth-aware image-space fog synthesis based on an atmospheric scattering model (Sakaridis et al., 2018), our renderer uses volumetric simulation and ray marching to generate spatially and temporally varying visual occlusion. The simulated smoke varies with position, time, and camera viewpoint, producing spatially non-uniform degradation in the robot's RGB observations. SmokeLab integrates comprehensive configurable parameters to flexibly adjust smoke color, shape, density and other core properties, enabling the generation of diverse visual interference patterns including different-density smoke, mixed-color smoke, steam, mist and dust.

Environment Setup. SmokeLab constructs diverse robot manipulation scenarios and interactive object assets. Figure 5 visualizes the ten operational scenes across five scenario classes; further details are provided in Appendix C. All scene assets, physical properties, semantic labels and rendering parameters are uniformly stored in Universal Scene Description (USD) files for standardized asset management and flexible scheduling. We adopt cuRobo (Sundaralingam et al., 2026) for motion

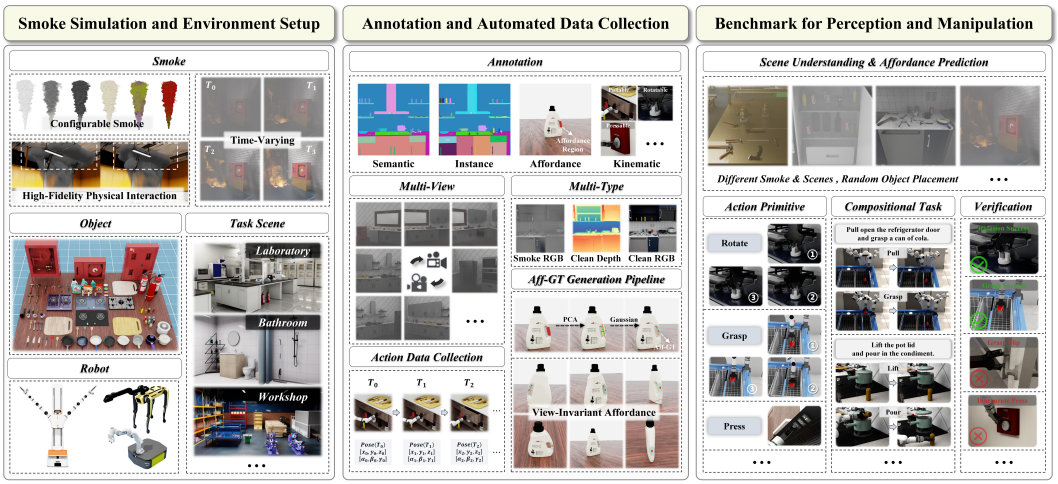


Figure 4: **Main modules of SmokeLab**. The platform provides configurable volumetric smoke and task scenes, automated multi-view data collection with diverse annotations, and standardized primitive and compositional tasks with execution verification.

generation in mobile manipulation. Each simulation episode of SmokeLab instantiates complete and independent simulation elements, including task scenes, robot embodiments, interactive object sets, camera configurations and customized smoke conditions, ensuring high reproducibility and configurability of all experimental trials across varied task configurations.

5.2 Annotation and Automated Data Collection

Annotation. We conduct fine-grained annotation for all manipulation objects in task scenarios, covering complete semantic categories and instance-level masks. In addition to basic visual labels, we further annotate the physical attributes of each object and precise affordance regions, providing multi-dimensional supervised information for visual perception and manipulation learning.

Automated Data Collection. Based on elaborate semantic and instance annotations, SmokeLab supports fully automated multi-view data collection for scene understanding tasks. We adopt the centroid, orientation, and axis extents estimated via PCA on affordance regions as the parameters of an anisotropic Gaussian distribution to generate affordance ground truth. This standardized generation pipeline produces reliable affordance labels for arbitrary camera viewpoints, enabling robots to learn robust view-invariant affordance representation for smoke-scene manipulation tasks.

5.3 Benchmark for Perception and Manipulation

SmokeLab establishes a hierarchical benchmark with four evaluation tracks to support a comprehensive study of robotic perception and manipulation under smoke. All four tracks use three smoke-density levels: Low for $\beta \in [0.1, 0.3)$, Medium for $\beta \in [0.3, 0.5)$, and High for $\beta \in [0.5, 5]$. The physical interpretation of β is provided in Appendix C.1. We construct dedicated datasets for perception evaluation tracks, containing 9072 images for scene understanding and 6048 images for affordance prediction, spanning diverse task scenarios.

The primitive execution track covers five fundamental robotic action primitives: grasp, pull, push, rotate, and press. The compositional manipulation track requires sequential execution of multiple primitive actions to complete long-horizon complex tasks. Detailed action evaluation metrics and task success verification mechanisms are elaborated in Appendix C.2. This hierarchical benchmark design enables fine-grained and systematic evaluation of robotic manipulation robustness under varying smoke interference and across diverse object interactions.

Table 1: Affordance prediction under smoke using SmokeScene-predicted masks.

Method	Low			Medium			High		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
AffordanceLLM	1.841	0.370	10.833	2.082	0.331	10.005	2.365	0.286	9.124
OS-AGDO	1.291	0.526	13.054	1.463	0.472	12.148	1.685	0.414	11.206
GLOVER++	1.243	0.523	12.449	1.397	0.468	11.532	1.572	0.409	10.617
SmokeAfford	**0.833**	**0.592**	**14.526**	**0.875**	**0.578**	**14.310**	**0.912**	**0.564**	**14.090**

Table 2: Average manipulation success rate (SR, %) for action primitives and compositional tasks. "Pri." and "Com." denote primitive and compositional tasks, respectively.

Method	Low		Medium		High		Real-world	
	Pri.	Com.	Pri.	Com.	Pri.	Com.	Pri.	Com.
$\pi_{0.5}$	46.4	24.5	26.8	11.9	9.8	4.6	14.9	4.5
OpenVLA-OFT	41.6	20.5	22.0	9.5	7.4	3.1	11.2	3.0
StableVLA	53.4	29.4	31.8	16.8	14.6	8.1	20.2	6.9
CRT-VLA	61.2	34.3	36.6	21.7	20.0	10.6	24.9	9.7
DehazeNet + DP	32.4	15.1	16.3	6.9	5.1	2.1	8.4	2.0
DehazeFormer + DP	34.7	16.8	18.1	7.7	5.8	2.4	9.3	2.3
SmokeManip	**71.1**	**53.3**	**60.0**	**46.7**	**44.5**	**33.3**	**44.5**	**28.9**

6 EXPERIMENT

6.1 BASELINES AND ABLATIONS SETUP

Affordance prediction baselines. We compare SmokeAfford with AffordanceLLM (Qian et al., 2024), OS-AGDO (Jia et al., 2025), and GLOVER++ (Ma et al., 2025), retrained on SmokeLab using their official weights and standard implementations. Each method is evaluated with SmokeScene-predicted and ground-truth target masks using the Kullback–Leibler divergence (KLD↓), Similarity (SIM↑), and Normalized Scanpath Saliency (NSS↑), following prior work (Qian et al., 2024).

Robotic manipulation baselines. We compare SmokeManip with general VLA policies: $\pi_{0.5}$ (Black et al., 2025) and OpenVLA-OFT (Kim et al., 2025); degradation-aware VLA methods: StableVLA (Fu et al., 2026) and CRT-VLA (Guarnizo Orjuela et al., 2026); and dehazing-based imitation-learning pipelines using DehazeNet (Cai et al., 2016)/DehazeFormer (Song et al., 2023) with Diffusion Policy (Chi et al., 2023). All baselines are adapted to SmokeLab using standard implementations; details are provided in Appendix D. We evaluate action primitives and compositional tasks with success rate (SR).

SmokeManip ablations. We compare the full model with six variants: **w/o SS** replaces SmokeScene with MaskDINO; **w/o SA** replaces SmokeAfford with GLOVER++; **w/o MC** removes target-mask conditioning; **w/o DA** replaces the smoke-adapted depth estimator with the original model for lifting; **w/o PG** removes pose generation and uses a fixed object-normal interaction pose; and **w/o CR** removes closed-loop recovery, executing once without retry or re-observation.

6.2 COMPARISON WITH BASELINES

Comparison with affordance prediction baselines. Table 1 compares affordance prediction methods using masks predicted by SmokeScene (the comparison between SmokeScene and baselines is reported in Appendix E). SmokeAfford consistently achieves the best KLD, SIM, and NSS across all smoke densities, with its advantage increasing under denser smoke. Its object-conditioned design preserves manipulation-relevant cues when RGB appearance is degraded. Results with ground-truth target masks in Appendix F further demonstrate robustness to mask quality.

Comparison with robotic manipulation baselines. Table 2 compares manipulation success rates across simulated smoke densities and real-world conditions. SmokeManip consistently outperforms

Table 3: Ablation study of the SmokeManip execution framework. Primitive and compositional success rates are reported at medium smoke density.

	SmokeManip	**w/o SS**	**w/o SA**	**w/o MC**	**w/o DA**	**w/o PG**	**w/o CR**
Primitive SR	**60.0**	44.5	37.8	28.9	20.0	17.8	42.2
Compositional SR	**46.7**	35.5	28.9	22.2	15.6	13.3	31.1

all baselines across both primitive and compositional manipulation tasks under low-, medium-, and high-density smoke. Both end-to-end VLA policies and dehazing-based imitation-learning pipelines struggle with the complex visual degradation induced by smoke. While all methods degrade as smoke density increases, SmokeManip retains a clear advantage by combining smoke-robust target localization, affordance prediction, and closed-loop execution.

6.3 Ablation Studies

Table 3 evaluates the SmokeManip execution components at medium smoke density. Replacing SmokeScene or SmokeAfford consistently reduces both primitive and compositional success rates, showing that reliable target localization and contact prediction are both necessary under smoke. Removing mask conditioning causes a further decline, while replacing the smoke-adapted depth estimator or removing pose generation produces the largest losses, highlighting the importance of accurate candidate lifting and executable pose selection. Disabling closed-loop recovery also lowers performance, confirming that retry and re-observation complement the perception and execution modules. Together, these results demonstrate the soundness of the framework design and the effectiveness of its components. Appendix G further validates the effectiveness of the perception components.

6.4 Real-World Evaluation

Figure 6 shows representative real-world evaluation scenes. To demonstrate real-world manipulation performance, we deployed our method on a RealMan robot and evaluated its performance. The real-world evaluation uses the same SmokeScene, SmokeAfford, and SmokeManip pipeline as simulation, without ground-truth masks or active depth at deployment, under medium smoke density. Table 2 reports success rates for action primitives and compositional tasks. SmokeManip consistently outperforms all baselines in both task categories, showing that its global-to-local perception and closed-loop execution remain effective under physical smoke. These experiments cover non-uniform smoke boundaries, and sensor and control noise not fully captured in simulation. More evaluation task details are provided in Appendix H.

7 Conclusion and Limitation

In this paper, we tackle robotic manipulation in smoke scenes and introduce SmokeManip, a perception-to-manipulation framework, together with SmokeLab, a simulation platform for evaluation. SmokeManip combines smoke-robust scene understanding, object-centric affordance prediction, and primitive-based closed-loop execution without active depth sensing at deployment. SmokeLab provides dynamic volumetric smoke, automated data collection, and unified protocols for robotic perception and manipulation. Experiments across smoke densities and physical scenes demonstrate the effectiveness of the framework for perception, action primitives, and compositional tasks. Despite the realism of SmokeLab's simulation, a gap between simulation and the real world remains unavoidable. Bridging this sim-to-real gap is therefore a key direction for future work.

AI USE STATEMENT

AI-based tools were used as assistants during the research and writing process. They supported the implementation and debugging of model code, drafting and language polishing of the manuscript, and the creation of visualizations and statistical plots from experimentally obtained data. All research decisions, experimental design, implementation validation, data analysis, result interpretation, and final manuscript revisions were conducted and verified by the authors. The authors take full responsibility for the content of this manuscript and the reported results.

REPRODUCIBILITY STATEMENT

We will release SmokeLab together with the source code and weights of the models.

REFERENCES

Chanh D. Tr. Nguyen, Jihyuk Park, Kyeong-Yong Cho, Kyung-Soo Kim, and Soohyun Kim. Novel descattering approach for stereo vision in dense suspended scatterer environments. *Sensors*, 17 (6):1425, June 2017. doi: 10.3390/s17061425.

Yuki Fujimura, Masaaki Iiyama, Atsushi Hashimoto, and Michihiko Minoh. Photometric stereo in participating media using an analytical solution for shape-dependent forward scatter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):708–719, 2020a. doi: 10.1109/TPAMI.2018.2889088.

Yuki Fujimura, Motoharu Sonogashira, and Masaaki Iiyama. Simultaneous estimation of object region and depth in participating media using a tof camera. *IEICE Transactions on Information and Systems*, E103.D(3):660–673, March 2020b. doi: 10.1587/transinf.2019EDP7219.

Devansh Dhrafani, Yifei Liu, Andrew Jong, Ukcheol Shin, Yao He, Tyler Harp, Yaoyu Hu, Jean Oh, and Sebastian Scherer. FIReStereo: Forest InfraRed stereo dataset for UAS depth perception in visually degraded environments. *IEEE Robotics and Automation Letters*, 10(4):3302–3309, 2025. doi: 10.1109/LRA.2025.3536278.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 2018.

Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding. *International Journal of Computer Vision*, 128(5):1182–1204, 2020. doi: 10.1007/s11263-019-01182-4.

Haifeng Zhong, Fan Tang, Hyung Jin Chang, Xingyu Zhu, and Yixing Gao. DarkSeg: Infrared-driven semantic segmentation for garment grasping detection in low-light conditions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 17680–17687, 2025. doi: 10.1109/IROS60139.2025.11247198.

Xingyu Zhu, Haifeng Zhong, Yan Wu, Shan Luo, and Yixing Gao. Guiding robotic cloth grasping in darkness: Infrared semantic segmentation and grasping position selection. *IEEE Robotics and Automation Letters*, 11(1):546–553, 2026a. doi: 10.1109/LRA.2025.3632718.

Mingdi Niu, Zhenyu Lu, Lu Chen, Jing Yang, and Chenguang Yang. VERGNet: Visual enhancement guided robotic grasp detection under low-light condition. *IEEE Robotics and Automation Letters*, 8(12):8541–8548, 2023. doi: 10.1109/LRA.2023.3330664.

Kun Zhang, Qinghua Li, Kaiyue Liu, Mengyao Zhang, Xuyang Wang, and Chao Feng. SPANet—sparse convolutional pyramid attention network for grasping detection in low-light conditions. In *China Automation Congress*, pages 5674–5679, 2023. doi: 10.1109/CAC59555.2023.10451820.

Haifeng Zhong, Wenshuo Han, Zhouyu Wang, Runyang Feng, Fan Tang, Tong-Yee Lee, Zipei Fan, Ruihai Wu, Yuran Wang, Hao Dong, Hechang Chen, Hyung Jin Chang, and Yixing Gao. GraspALL: Adaptive structural compensation from illumination variation for robotic garment grasping in any low-light conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6641, 2026.

Yiyang Fu, Chubin Zhang, Shukai Gong, Yufan Deng, Kaiwei Sun, Qiyang Min, Qibin Hou, Yansong Tang, Jianan Wang, and Daquan Zhou. StableVLA: Towards robust vision-language-action models without extra data. In *International Conference on Machine Learning*, 2026.

Daniel Yezid Guarnizo Orjuela, Leonardo Scappatura, Veronica Di Gennaro, Riccardo Andrea Izzo, Gianluca Bardaro, and Matteo Matteucci. Improving robustness of vision-language-action models by restoring corrupted visual inputs, 2026.

Jiajun Zhai, Hao Shi, Shangwei Guo, Kailun Yang, and Kaiwei Wang. E-VLA: Event-augmented vision-language-action model for dark and blurred scenes. In *European Conference on Computer Vision*, 2026.

Ganlin Yang, Zhangzheng Tu, Yuqiang Yang, Sitong Mao, Junyi Dong, Tianxing Chen, Jiaqi Peng, Jing Xiong, Jiafei Cao, Jifeng Dai, Wengang Zhou, Yao Mu, and Tai Wang. EventVLA: Event-driven visual evidence memory for long-horizon vision-language-action policies, 2026.

Liang Liao, Wenyi Chen, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Unsupervised foggy scene understanding via self spatial-temporal label diffusion. *IEEE Transactions on Image Processing*, 31:3525–3540, 2022. doi: 10.1109/TIP.2022.3172208.

Bo Liu, Si-Bao Chen, Jia-Xin Wang, Jin Tang, and Bin Luo. An oriented object detector for hazy remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–11, 2024. doi: 10.1109/TGRS.2024.3451492.

Zhenbing Liu, Tianle Fang, Haoxiang Lu, Weidong Zhang, and Rushi Lan. MASFNet: Multiscale adaptive sampling fusion network for object detection in adverse weather. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025a. doi: 10.1109/TGRS.2025.3558541.

Changfeng Feng, Zhenyuan Chen, Xiang Li, Chunping Wang, Yimian Dai, Qiang Fu, Jian Yang, and Ming-Ming Cheng. HazyDet: Open-source benchmark for drone-view object detection with depth cues in hazy scenes. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5006414, 2026. doi: 10.1109/TGRS.2026.3701819.

Sohyun Lee, Taeyoung Son, and Suha Kwak. FIFO: Learning fog-invariant features for foggy scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18911–18921, 2022.

Weifeng Liu, Jian Pang, Bingfeng Zhang, Jin Wang, Baodi Liu, and Dapeng Tao. See degraded objects: A physics-guided approach for object detection in adverse environments. *IEEE Transactions on Image Processing*, 34:2198–2212, 2025b. doi: 10.1109/TIP.2025.3551533.

Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01835.

Mingjia Li, Binhui Xie, Shuang Li, Chi Harold Liu, and Xinjing Cheng. VBLC: Visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Donggon Jang, Sunhyeok Lee, Gyuwon Choi, Yejin Lee, Sanghyeok Son, and Dae Shik Kim. Energy-based domain adaptation without intermediate domain dataset for foggy scene segmentation. *IEEE Transactions on Image Processing*, 33:6143–6157, 2024. doi: 10.1109/TIP.2024.3483566.

Shiqin Wang, Haoyang Chen, Huaizhou Huang, Yinkan He, Dongfang Sun, Xiaoqing Chen, Xingyu Liu, Zheng Wang, and Kaiyan Zhao. Heuristic self-paced learning for domain adaptive semantic segmentation under adverse conditions, 2026.

Wanjun Jia, Fan Yang, Mengfei Duan, Xianchi Chen, Yinxi Wang, Yiming Jiang, Wenrui Chen, Kailun Yang, and Zhiyong Li. One-shot affordance grounding of deformable objects in egocentric organizing scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 21243–21250, 2025. doi: 10.1109/IROS60139.2025.11246431.

Mengzhen Liu, Enshen Zhou, Cheng Chi, Yi Han, Shanyu Rong, Liming Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. SaPaVe: Towards active perception and manipulation in vision-language action models for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37164–37174, 2026.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Erwin Coumans and Yunfei Bai. PyBullet, a python module for physics simulation for games, robotics and machine learning, 2021.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

NVIDIA. Isaac Sim, 2026. Version 6.0.1.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark and learning environment. *IEEE Robotics and Automation Letters*, 2020.

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Proceedings of the Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 80–93, 2023b.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. In *Proceedings of Robotics: Science and Systems*, 2024.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. RoboCasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, 2026.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1094–1100, 2020.

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, Xiaodi Yuan, Pengwei Xie, Zhiao Huang, Rui Chen, and Hao Su. ManiSkill2: A unified benchmark for generalizable manipulation skills. In *International Conference on Learning Representations*, 2023.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27649–27660, 2025. doi: 10.1109/CVPR52734.2025.02575.

Zhou Xian, Bo Zhu, Zhenjia Xu, Hsiao-Yu Tung, Antonio Torralba, Katerina Fragkiadaki, and Chuang Gan. FluidLab: A differentiable environment for benchmarking complex fluid manipulation. In *International Conference on Learning Representations*, 2023.

Yian Wang, Juntian Zheng, Zhehuan Chen, Zhou Xian, Gu Zhang, Chao Liu, and Chuang Gan. Thinshell object manipulations with differentiable physics simulations. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 42647–42667, 2024.

Yuran Wang, Ruihai Wu, Yue Chen, Jiarui Wang, Jiaqi Liang, Ziyu Zhu, Haoran Geng, Jitendra Malik, Pieter Abbeel, and Hao Dong. DexGarmentLab: Dexterous garment manipulation environment with generalizable policy. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Junyi Cao, Yian Wang, Ziyan Xiong, Chunru Lin, Zhehuan Chen, and Chuang Gan. DLO-Lab: Benchmarking deformable linear object manipulations with differentiable physics. In *International Conference on Machine Learning*, 2026.

Xingyu Zhu, Wenshuo Han, Zhouyu Wang, Yuran Wang, Ruihai Wu, Hao Dong, Fan Tang, Hechang Chen, Hyung Jin Chang, and Yixing Gao. FlatLab: A unified methodology framework and simulation-based benchmark for robotic manipulation of flat objects, 2026b.

Rui Li, Zixuan Hu, Wenxi Qu, Jinouwen Zhang, Zhenfei Yin, Sha Zhang, Xuantuo Huang, Hanqing Wang, Tai Wang, Jiangmiao Pang, Wanli Ouyang, Lei Bai, Wangmeng Zuo, Ling-Yu Duan, Dongzhan Zhou, and Shixiang Tang. LabUtopia: High-fidelity simulation and hierarchical benchmark for scientific embodied agents. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-1158.

Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, Xiaokang Yang, Xuelong Li, Hongyuan Zhang, Yao Mu, and Ping Luo. AutoBio: A simulation and benchmark for robotic automation in digital biology laboratory. In *International Conference on Learning Representations*, 2026.

Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023. doi: 10.1109/TRO.2023.3281153.

Alexandros Paraschos, Christian Daniel, Jan R. Peters, and Gerhard Neumann. Probabilistic movement primitives. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Balakumar Sundaralingam, Adithyavairavan Murali, and Stan Birchfield. cuRoboV2: Dynamics-aware motion generation with depth-fused distance fields for high-dof robots, 2026.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. AffordanceLLM: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 7587–7597, June 2024.

Teli Ma, Jia Zheng, Zifan Wang, Ziyao Gao, Jiaming Zhou, and Junwei Liang. GLOVER++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation, 2025.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Richter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In Joseph Lim, Shuran Song, and Hae-Won Park, editors, *Proceedings of the Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 17–40. PMLR, 2025.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.017.

Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. DehazeNet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. doi: 10.1109/TIP.2016.2598681.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. doi: 10.1109/TIP.2023.3256763.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026.

Srinivasa G. Narasimhan and Shree K. Nayar. Vision and the atmosphere. *International Journal of Computer Vision*, 48(3):233–254, 2002. doi: 10.1023/A:1016328200723.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023c.

Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. OMG-Seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tommie Kerssies, Niccolò Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your ViT is secretly an image segmentation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

APPENDIX CONTENTS

A SmokeManip Implementation Details **16**

A.1 Candidate Representation and Scoring 16

A.2 Action Primitive Library 16

A.3 Perception Implementation Details 16

A.4 Fine-tuned Monocular Depth for Candidate Lifting 17

B SmokeScene and SmokeAfford Training Details **17**

C Details of SmokeLab **19**

C.1 Smoke Density Quantification via Beta Value 19

C.2 Evaluation Criteria for Primitive and Compositional Tasks 20

D Baseline Adaptation Details **21**

E Scene Understanding Baselines and Results **21**

F Affordance Results with Ground-Truth Masks **21**

G Perception Ablation Studies **22**

H Real-World Success Rates **22**

I Details of Real-World Evaluation **22**

A SmokeManip Implementation Details

All model training and DepthAnythingv2 fine-tuning were conducted on NVIDIA RTX 4090 GPUs.

A.1 Candidate Representation and Scoring

Each contact or grasp candidate is represented by $\mathbf{g}_k = (R_k, \mathbf{t}_k) \in SE(3)$, where $R_k \in SO(3)$ is its orientation and $\mathbf{t}_k \in \mathbb{R}^3$ is its position. Candidates are ranked by

$$S(\mathbf{g}_k) = \lambda_a S_a(\mathbf{g}_k) + \lambda_p S_p(\mathbf{g}_k), \tag{6}$$

where $S_a(\mathbf{g}_k)$ is the affordance heatmap response, $S_p(\mathbf{g}_k)$ is the pose-generator confidence, and λ_a and λ_p are weighting coefficients.

A.2 Action Primitive Library

SmokeManip employs a lightweight action primitive library for action execution rather than learning monolithic trajectory policies. This library generates constrained Cartesian motion templates, while language grounding, gripper-contact scheduling, feasibility constraints and recovery logic are handled by the SmokeManip execution wrapper.

Template collection and fitting. For each action verb supported by the library, multiple clean scripted reference motions are collected in simulation to learn trajectory-weight distributions. These demonstrations are free of smoke corruption and are not used to train SmokeScene or SmokeAfford; they encode canonical motion patterns for task execution. At runtime, primitive distributions are conditioned on the end-effector start pose and contact via-points obtained from affordance-guided pose generation. The mean of conditioned distributions yields nominal Cartesian reference trajectories. Gripper or contact-state changes are stored as synchronized schedules. Each primitive additionally stores feasibility constraints and object-state success predicates. The resulting library generalizes across diverse object geometries and scene layouts and supports easy extension to additional action types.

Primitive library. The library contains five action primitives: *grasp*, *pull*, *push*, *rotate*, and *press*. Parsed task verbs index corresponding primitives; target categories and object states parameterize motion ranges such as pushing displacement or rotation axis. Contact poses derived from object-centric affordances anchor each primitive for trajectory instantiation.

Object-state predicates. Primitive completion is judged by task-aligned object-state predicates following standard manipulation-benchmark evaluation criteria instead of raw trajectory termination signals. For instance, a grasp succeeds when the target remains held after lifting; push succeeds upon crossing displacement thresholds; rotate succeeds once object orientation exceeds an angular threshold. These predicates supply failure signals driving the hierarchical closed-loop recovery.

Retry workflow. Upon receiving failure signals from object-state predicates, the system activates the hierarchical recovery pipeline elaborated in the main text.

A.3 Perception Implementation Details

Mask refinement. The residual target-mask update is

$$\widehat{\mathcal{M}}^{+} = \widehat{\mathcal{M}} + \alpha_{\mathrm{M}} \left[\mathcal{R}_{\mathrm{M}} \left(F^{\mathrm{M}}, \widehat{\mathcal{M}}, E^{\mathrm{D}} \right) - \widehat{\mathcal{M}} \right]. \tag{7}$$

Here, $\widehat{\mathcal{M}}$ and $\widehat{\mathcal{M}}^{+}$ are the input and refined soft masks, respectively. The refiner $\mathcal{R}_{\mathrm{M}}$ is conditioned on F^{M}, $\widehat{\mathcal{M}}$, and E^{D}, while α_{M} controls the residual update magnitude.

Object-centric attention and affordance output. The category-conditioned object tokens attend to the fused spatial features as

$$\mathbf{Z}^{\mathrm{O}} = \mathcal{T}_{\mathrm{O}}\big(\text{Query} = T^{\mathrm{O}} + \mathbf{e},\ \text{Key=Value} = \mathrm{Concat}\left(\widetilde{F}^{\mathrm{P}}, F^{\mathrm{M}}, F^{\mathrm{D}}\right)\big). \tag{8}$$

The affordance decoder then maps the conditioned feature, object representation, and refined mask to the action-specific affordance heatmap:

$$\widehat{\mathbf{H}} = \mathcal{D}_{\mathrm{A}}\Big(F^{\mathrm{A}}, \mathbf{Z}^{\mathrm{O}}, \widehat{\mathcal{M}}^{+} \Big). \tag{9}$$

Here, $\mathcal{D}_{\mathrm{A}}$ denotes the affordance decoder.

Geometry supervision. The geometry term in Eq. 17 is applied to selected backbone and mask-decoding stages. Depth-derived structural targets are used as privileged supervision during training, and no depth input is used by the scene parser at deployment. Loss weights and severity gates are selected on the validation split and reported with the released training configuration.

Affordance regularization. The explicit target-mask-containment, peak-sharpening, geometry, and uncertainty terms in Eq. 20 are expanded in Appendix B.

Affordance model. SmokeAfford receives a smoke RGB image and a target-object mask $\widehat{\mathcal{M}}$. Predicted masks are converted to soft masks and resized to the affordance feature resolution before mask encoding. These operations align feature resolutions and improve numerical stability.

Hyperparameters. Concrete choices such as query count, feature resolution, heatmap Gaussian radius, loss weights, mask-curriculum schedule, and optimizer settings are reported in the released configuration files when they affect reproducibility.

A.4 Fine-tuned Monocular Depth for Candidate Lifting

The smoke-adapted depth estimator lifts ranked 2D affordance candidates into 3D grasp or contact hypotheses. In our implementation, $\widehat{D}(\cdot)$ is a DepthAnythingv2 monocular depth estimator fine-tuned on SmokeLab RGB–depth pairs. It is used only for action execution: SmokeScene and SmokeAfford do not require active depth input or ground-truth masks during deployment. The off-the-shelf DepthAnythingv2 model produces unreliable depth estimates under smoke scattering, so fine-tuning is required to meet the precision requirements of robotic manipulation.

Dense smoke changes the image statistics seen by an off-the-shelf monocular depth model. In global dense smoke, depth quality can degrade because the target boundary and background contrast are suppressed. SmokeManip therefore uses the global RGB observation only for coarse approach and triggers a local re-observation before 2D-to-3D lifting. In the local RGB observation, the target occupies more pixels and the smoke field is less ambiguous inside the target-object mask, so the smoke-adapted depth estimator provides sufficient local depth and surface-normal accuracy for candidate ranking and robot-frame pose construction. Table 3 includes an execution ablation that replaces this estimator with an off-the-shelf monocular depth model; the drop quantifies the dependence of 2D-to-3D lifting on the smoke-adapted depth estimator.

B SmokeScene and SmokeAfford Training Details

Scene parser objective. The base query objective is decomposed into semantic and instance terms. The semantic term is

$$\mathcal{L}_{\mathrm{sem}} = -\lambda_{\mathrm{cls}} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\alpha y_i (1 - p_i)^{\gamma} \log p_i + (1 - \alpha)(1 - y_i) p_i^{\gamma} \log(1 - p_i) \right]. \tag{10}$$

where ℓ_i and y_i are a query–class logit and its binary target, $p_i = \mathrm{sigmoid}(\ell_i)$, $\gamma = 2$, and α balances positive and negative classes. The instance term combines the mask and Dice objectives:

$$\mathcal{L}_{\mathrm{ins}} = \lambda_{\mathrm{mask}} \mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{dice}} \mathcal{L}_{\mathrm{dice}}. \tag{11}$$

For matched mask points, the focal-style mask term is

$$\mathcal{L}_{\mathrm{mask}} = -\frac{1}{|\mathcal{P}|} \sum_{u \in \mathcal{P}} \left[\alpha m_u (1 - q_u)^{\gamma} \log q_u + (1 - \alpha)(1 - m_u) q_u^{\gamma} \log(1 - q_u) \right]. \tag{12}$$

The Dice term is

$$\mathcal{L}_{\mathrm{dice}} = 1 - \frac{2 \sum_{u \in \mathcal{P}} q_u m_u + 1}{\sum_{u \in \mathcal{P}} q_u + \sum_{u \in \mathcal{P}} m_u + 1}. \tag{13}$$

Here a_u and m_u are the sampled mask logit and target at point u, and $q_u = \mathrm{sigmoid}(a_u)$.

Geometry supervision. The local smoke-severity weight used in the geometry losses is

$$\omega(p) = \mathrm{Resize}\left(\exp\left(\gamma_{\mathrm{sev}} \frac{1 - \exp\big(-\beta D^{\mathrm{C}}(p)\big)}{\max_{q \in \Omega}(1 - \exp(-\beta D^{\mathrm{C}}(q))) + \epsilon} \right) \right). \tag{14}$$

Here, $D^{\mathrm{C}} \in [0,1]^{H\times W}$ is the clean-depth map, p is the current pixel, q indexes spatial normalization, and Ω is the image domain. The scalar γ_{sev} controls the strength of severity reweighting, while β controls its sensitivity to clean depth. The weight map is resized at each feature level to obtain ω_ℓ and ω_k for the encoder and decoder losses, respectively.

The encoder geometry loss is

$$\mathcal{L}_{\text{enc-geo}} = \frac{1}{|\mathcal{K}_{\text{enc}}|} \sum_{\ell\in\mathcal{K}_{\text{enc}}} \left[\mathcal{L}_{\cos}(F_\ell^{\text{Enc}}, F^{\text{T}}, \omega_\ell) + \lambda_{\text{sobel}}\mathcal{L}_{\text{sobel}}^{\text{enc}}(F_\ell^{\text{Enc}}, F^{\text{T}}, \omega_\ell)\right], \tag{15}$$

where $\mathcal{K}_{\text{enc}} = \{1,2\}$ indexes the first two encoder feature levels. The decoder geometry loss is

$$\mathcal{L}_{\text{dec-geo}} = \frac{1}{|\mathcal{K}_{\text{dec}}|} \sum_{k\in\mathcal{K}_{\text{dec}}} \mathcal{L}_{\text{sobel}}^{\text{dec}}\big(F_k^{\text{Dec}}, F^{\text{T}}, \omega_k\big), \tag{16}$$

where $\mathcal{K}_{\text{dec}} = \{1,2,3,\text{mask}\}$ indexes the three multi-scale pixel-decoder features and the mask feature. The total SmokeScene objective is

$$\mathcal{L}_{\text{scene}} = \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{ins}} + \alpha_{\text{geo}}(t)\left(\lambda_{\text{enc}}\mathcal{L}_{\text{enc-geo}} + \lambda_{\text{dec}}\mathcal{L}_{\text{dec-geo}}\right). \tag{17}$$

Here, $\alpha_{\text{geo}}(t)$ applies warm-up and late decay to the complete geometry objective, while the severity weight maps distribute that supervision spatially within each feature level.

For each encoder level, let $\overline{F}^{\text{Enc}} = \text{Norm}(\mathcal{P}(F^{\text{Enc}}))$ and $\overline{F}^{\text{T}} = \text{Norm}(\text{Resize}(F^{\text{T}}))$ denote the aligned encoder and depth-VFM teacher features, respectively, where $\mathcal{P}$ is a learned channel projection. The cosine term in Eq. 15 is

$$\mathcal{L}_{\cos}(F^{\text{Enc}}, F^{\text{T}}, \omega) = \frac{\sum_{x\in\Omega}\omega_x\left(1 - \frac{\left\langle \overline{F}_x^{\text{Enc}}, \overline{F}_x^{\text{T}}\right\rangle}{\|\overline{F}_x^{\text{Enc}}\|_2\|\overline{F}_x^{\text{T}}\|_2+\epsilon}\right)}{\sum_{x\in\Omega}\omega_x}, \tag{18}$$

where Ω indexes spatial positions and ω_x is the smoke-severity weight. Thus, $\mathcal{L}_{\cos}$ measures the severity-weighted cosine discrepancy between normalized projected encoder and teacher features. The encoder Sobel term compares the horizontal and vertical gradients of the same aligned features channel-wise:

$$\mathcal{L}_{\text{sobel}}^{\text{enc}}(F^{\text{Enc}}, F^{\text{T}}, \omega) = \frac{\sum_{x\in\Omega}\omega_x\ \text{Mean}_c\left[\sum_{a\in\{x,y\}}\rho\Big(S_a(\overline{F}^{\text{Enc}}) - S_a(\overline{F}^{\text{T}})\Big)_x\right]}{\sum_{x\in\Omega}\omega_x},$$

Here $\mathcal{L}_{\text{sobel}}^{\text{enc}}$ computes channel-wise horizontal and vertical Sobel-gradient matching via severity-weighted Smooth-L1. Here S_x and S_y are the horizontal and vertical Sobel operators, a indexes these directions, and ρ denotes Smooth-L1.

For decoder level k, let $F_k^{\text{T}} = \text{Resize}_k(F^{\text{T}})$. The normalized structural response and its weighted loss are defined jointly as

$$\begin{aligned}\mathcal{E}(F) &= \text{Norm}(\|\nabla_{\text{Sobel}}\,\text{Mean}_c(F)\|_2),\\ \mathcal{L}_{\text{sobel}}^{\text{dec}}(F^{\text{Dec}}, F^{\text{T}}, \omega) &= \frac{\sum_{x\in\Omega}\omega_x\,\rho\big(\mathcal{E}(F^{\text{Dec}})_x - \mathcal{E}(F_k^{\text{T}})_x\big)}{\sum_{x\in\Omega}\omega_x}.\end{aligned} \tag{19}$$

where $\text{Mean}_c(\cdot)$ reduces each feature to a single-channel map before gradient extraction. The decoder loss computes normalized Sobel-response matching at each decoder resolution via severity-weighted Smooth-L1.

Affordance objective. The total affordance objective is

$$\mathcal{L}_{\text{afford}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{kld}}\mathcal{L}_{\text{kld}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{aux}}. \tag{20}$$

For condition s, let $\widehat{Y}_{n,x}^{(s)} = \sigma(z_{n,x}^{(s)})$ and $P_{n,x}^{(s)} = \text{softmax}_x(z_{n,x}^{(s)})$. The normalized ground-truth distribution is $\widetilde{Y}_{n,x} = Y_{n,x}/\sum_{q\in\Omega}Y_{n,q}$, and the heatmap-dependent weight is $w_{n,x} = 1 + \lambda_{\text{pos}}Y_{n,x}/\max_{q\in\Omega}Y_{n,q}$. The weighted pixel-wise loss is

$$\mathcal{L}_{\text{bce}} = -\frac{1}{B|\Omega|}\sum_{n=1}^{B}\sum_{x\in\Omega} w_{n,x}\left[Y_{n,x}\log\widehat{Y}_{n,x}^{(s)} + (1-Y_{n,x})\log(1-\widehat{Y}_{n,x}^{(s)})\right]. \tag{21}$$

The distribution-alignment loss is

$$\mathcal{L}_{\mathrm{kld}} = \frac{1}{B}\sum_{n=1}^{B}\sum_{x\in\Omega}\widetilde{Y}_{n,x}\log\frac{\widetilde{Y}_{n,x}}{P^{(s)}_{n,x}+\epsilon}. \tag{22}$$

The mask-perturbation consistency term combines distribution- and pixel-level agreement. The original-mask prediction is treated as a stop-gradient reference:

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{B}\sum_{n=1}^{B}\sum_{x\in\Omega}\mathrm{sg}(P^{(0)}_{n,x})\log\frac{\mathrm{sg}(P^{(0)}_{n,x})}{P^{(1)}_{n,x}+\epsilon} + \frac{1}{B|\Omega|}\left\|\widehat{\mathbf{Y}}^{(1)} - \mathrm{sg}(\widehat{\mathbf{Y}}^{(0)})\right\|_1. \tag{23}$$

Auxiliary affordance constraints. The auxiliary term in Eq. 20 is

$$\mathcal{L}_{\mathrm{aux}} = \lambda_{\mathrm{mse}}\mathcal{L}^{w}_{\mathrm{MSE}} + \lambda_{\mathrm{coord}}\mathcal{L}_{\mathrm{coord}} + \lambda_{\mathrm{iop}}\mathcal{L}_{\mathrm{iop}} + \lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}. \tag{24}$$

Here, $\mathcal{L}^{w}_{\mathrm{MSE}}$ uses the same heatmap-dependent weights as the weighted BCE, and $\mathcal{L}_{\mathrm{coord}}$ is a Smooth-L1 loss on soft-argmax coordinates. Specifically,

$$\mathcal{L}^{w}_{\mathrm{MSE}} = \frac{\sum_{n,x} w_{n,x}(\widehat{Y}^{(s)}_{n,x} - Y_{n,x})^2}{\sum_{n,x} w_{n,x}}, \qquad \mathcal{L}_{\mathrm{coord}} = \mathrm{SmoothL1}\Big(\mathrm{SoftArgmax}(\mathbf{z}^{(s)}), \mathrm{Coord}(\mathbf{Y})\Big). \tag{25}$$

The inside/outside/peak term is

$$\mathcal{L}_{\mathrm{iop}} = -\frac{1}{B}\sum_{n=1}^{B}\log I_n + \frac{1}{B}\sum_{n=1}^{B}O_n - \lambda_{\mathrm{peak}}\frac{1}{B}\sum_{n=1}^{B}\log P_n, \tag{26}$$

where I_n and O_n are the predicted probability mass inside and outside the target-object mask for sample n, and P_n is the predicted probability at the annotated affordance peak. The first two terms impose target-mask containment, while the peak term sharpens the response around the annotated peak. The coefficient λ_{peak} controls peak sharpening. The geometry-edge term $\mathcal{L}_{\mathrm{edge}}$ supervises the predicted geometry edge map with a binary cross-entropy loss:

$$\mathcal{L}_{\mathrm{edge}} = -\frac{1}{|\Omega|}\sum_{x\in\Omega}\left[r^{\mathrm{gt}}_x\log\sigma(r_x) + (1-r^{\mathrm{gt}}_x)\log\left(1-\sigma(r_x)\right)\right], \tag{27}$$

where r_x and r^{gt}_x are the predicted geometry-edge logit and binary target at position x. The uncertainty penalty is defined separately:

$$\mathcal{L}_{\mathrm{unc}} = \frac{1}{|\Omega|}\sum_{x\in\Omega}\sigma(u_x)A^{\mathrm{gt}}_x. \tag{28}$$

where u_x is the uncertainty logit and A^{gt}_x is the annotated affordance map at position x. Thus $\mathcal{L}_{\mathrm{edge}}$ encourages boundary prediction, while $\mathcal{L}_{\mathrm{unc}}$ penalizes high predicted uncertainty near annotated affordance locations. The coefficient values are fixed by the released training configuration.

C Details of SmokeLab

SmokeLab curates and rearranges scene assets to construct ten operational scenes organized into five classes: bathrooms, kitchens, laboratories, stores, and workshops. All ten scenes are shown in Fig. 5. Each asset is annotated with empirically derived physical and semantic properties to support consistent simulation and evaluation. In addition to these curated scenes, SmokeLab supports the diverse kitchen assets provided by RoboCasa365 (Nasiriany et al., 2026).

C.1 Smoke Density Quantification via Beta Value

To achieve quantitative and controllable smoke interference, we adopt a calibrated beta value to determine the global smoke density of each simulation scenario. The beta value ranges from 0 to 5, corresponding to transparent smoke-free conditions and extremely dense smoke occlusion, respectively. The benchmark intervals are defined in Section 5.3. All smoke-related rendering parameters, including scattering coefficient, atmospheric light intensity and voxel density, are dynamically



Figure 5: **Visualization of various manipulation scenarios in SmokeLab**.

mapped to the beta value, ensuring stable and consistent smoke level division across all simulation scenes and tasks.

The use of β follows the atmospheric scattering model, in which the transmission of scene radiance along a path of length $d(x)$ is

$$t(x) = \exp(-\beta d(x)), \tag{29}$$

and the observed image is

$$I(x) = J(x)t(x) + A\,(1 - t(x)), \tag{30}$$

where $J(x)$ is the clear-scene radiance and A is atmospheric light. The product $\beta d(x)$ is the optical depth. For a fixed viewing distance, a larger β monotonically decreases transmission, increases the scattered airlight contribution, and reduces scene contrast. The extinction coefficient is therefore inversely related to visual range under the standard homogeneous atmosphere assumptions (Narasimhan and Nayar, 2002). In SmokeLab, β is a calibrated renderer control rather than a direct physical measurement; it provides a consistent quantitative proxy for smoke concentration and visibility across matched scenes.

C.2 Evaluation Criteria for Primitive and Compositional Tasks

Notably, precise metric depth is reserved only for geometric supervision during model training, while all evaluations use smoke-corrupted RGB images without active depth information.

C.2.1 Primitive Action Evaluation Metrics

Each fundamental robotic primitive action is evaluated via explicit object state predicates to eliminate subjective judgment:

- **Grasp**: The target object remains stably clamped without slipping or falling within the predefined holding duration after grasping execution.
- **Pull**: The target component produces effective displacement along the intended pulling direction without excessive offset or jitter.
- **Push**: The target object or component produces effective displacement along the intended pushing direction without excessive offset or jitter.
- **Rotate**: The object achieves the expected angular change within the error tolerance range, maintaining stable posture during rotation.
- **Press**: The target button or functional component generates valid state transition and maintains the pressed state after operation.

C.2.2 Compositional Task Success Mechanism

For compositional manipulation tasks with long-horizon sequential execution, we adopt a strict stepwise verification strategy for success judgment. A single trial is considered successful *only if* all state transitions of each primitive subtask are verified valid in strict chronological order. This mechanism explicitly distinguishes two types of task failures:

- **Contact-level failures**: Invalid execution of individual primitive actions, such as grasp slipping and failed pressing.

- **Cumulative execution errors**: Step-by-step state deviation and error accumulation in sequential multi-stage manipulation, even if each single primitive action is temporarily valid.

This fine-grained evaluation paradigm enables precise analysis of model failure modes in complex smoke-contaminated manipulation scenarios.

D Baseline Adaptation Details

To ensure a matched comparison in SmokeLab, all robotic manipulation baselines use the same camera observations, language instructions, robot action space, control frequency, initial states, and termination predicates as SmokeManip. We retain the standard implementation and training objective of each method, while replacing only robot-specific input/output adapters and fine-tuning on SmokeLab demonstrations. No baseline receives SmokeScene masks, SmokeAfford responses, smoke-adapted depth, or SmokeManip's closed-loop retry logic.

General VLA policies. For $\pi_{0.5}$ (Black et al., 2025), we preserve its hierarchical high-level semantic action prediction followed by low-level action-chunk generation, and map the SmokeLab language and robot state to the corresponding model interfaces. For OpenVLA-OFT (Kim et al., 2025), we retain the official parallel decoding, action chunking, continuous action representation, and L1 regression objective, while replacing the robot-specific action head and proprioceptive projection for SmokeLab. Both policies are fine-tuned on the SmokeLab RGB–language–action trajectories.

Degradation-aware VLA policies. For StableVLA (Fu et al., 2026), we retain the information-bottleneck adapter that replaces the vision–language projector and fine-tune its SmokeLab-compatible action interface on smoke observations. For CRT-VLA (Guarnizo Orjuela et al., 2026), we place the corruption-restoration transformer before the frozen VLA policy, train the restoration module with paired clean and smoke-degraded SmokeLab frames, and keep the downstream policy and action interface otherwise unchanged. These adaptations expose each method's stated robustness mechanism without adding SmokeManip-specific scene or affordance modules.

Dehazing plus imitation-learning policies. For the DehazeNet (Cai et al., 2016) and DehazeFormer (Song et al., 2023) pipelines, the dehazing network is trained on paired SmokeLab smoke/clean RGB frames and its restored image is passed to a Diffusion Policy (Chi et al., 2023). The policy is fine-tuned with the same SmokeLab action trajectories and robot action parameterization as the other baselines; no target-mask transfer, depth estimator, or retry controller is added.

For every atomic action primitive, we evaluate 10 independent episodes per method. Each compositional task is evaluated on 20 independent episodes. The same episode counts, initial configurations, language instructions, and action budgets are used for all baselines and SmokeManip.

E Scene Understanding Baselines and Results

We compare SmokeScene with the fog-adaptive methods CMAda (Ma et al., 2022) and FIFO (Lee et al., 2022), and the universal scene understanding models Mask2Former (Cheng et al., 2022), MaskDINO (Li et al., 2023c), OMG-Seg (Li et al., 2024), and EoMT (Kerssies et al., 2025). All baselines are retrained on SmokeLab using official pretrained weights and standard implementations. We report PQ, PQ^{th}, and Target Recall (TR), where TR is the proportion of task-relevant targets recovered with mask IoU ≥ 0.5.

Table 4 compares SmokeScene with smoke-adaptive and universal scene understanding methods across three smoke densities. SmokeScene achieves the best PQ, PQ^{th}, and TR at all density levels, while competing methods degrade more markedly as smoke density increases. In contrast, SmokeScene maintains stable performance by using geometric supervision to preserve structural cues under smoke degradation.

F Affordance Results with Ground-Truth Masks

Table 5 reports affordance prediction when the ground-truth target-object mask is provided.

Table 4: Scene understanding under smoke.

Method	Low			Medium			High		
	PQ↑	PQ^{th}↑	TR↑	PQ↑	PQ^{th}↑	TR↑	PQ↑	PQ^{th}↑	TR↑
FIFO	62.10	60.00	66.30	60.30	58.10	64.20	58.00	55.80	61.80
CMAda	61.20	59.10	65.20	59.40	57.20	63.10	57.10	54.90	60.70
OMG-Seg	64.14	61.50	69.40	61.34	58.70	66.60	55.27	53.00	60.80
EoMT	74.20	71.80	78.50	72.40	69.90	76.30	71.00	68.40	74.60
Mask2Former	75.04	72.50	81.20	73.25	70.70	79.10	71.50	68.90	76.80
MaskDINO	77.72	75.20	85.10	76.10	73.40	82.60	74.50	71.60	80.20
SmokeScene	**78.20**	**75.85**	**87.40**	**77.55**	**74.65**	**86.30**	**76.25**	**73.55**	**85.70**

Table 5: Manipulation affordance prediction under smoke using ground-truth target-object masks. KLD↓, SIM↑, and NSS↑ are reported at each smoke density.

Method	Low			Medium			High		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
AffordanceLLM	1.423	0.443	14.711	1.280	0.468	13.756	1.202	0.490	12.651
OS-AGDO	1.221	0.492	15.602	1.085	0.518	14.614	1.063	0.503	13.586
GLOVER++	3.710	0.216	5.809	3.181	0.240	6.166	3.173	0.191	4.658
SmokeAfford	**1.128**	**0.516**	**16.688**	**0.944**	**0.540**	**16.091**	**0.718**	**0.570**	**15.460**

G Perception Ablation Studies

Table 6 reports the SmokeScene and SmokeAfford component ablations omitted from the main-text execution study. The entries and metrics match those used in the main experimental analysis.

Removing object-mask refinement and encoding causes the largest degradation in SmokeAfford, showing that an accurate object-level spatial prior is essential for affordance prediction under smoke. Depth-guided feature modulation, object-centric representation modeling, and object-conditioned decoding also improve the three affordance metrics, confirming that each component contributes to the full model.

H Real-World Success Rates

We evaluate the complete SmokeManip pipeline on the same five action primitives used by SmokeLab: grasp, pull, push, rotate, and press. All real-world trials use medium smoke density and follow a three-round protocol with five episodes per round. The real-world evaluation further includes four compositional tasks: (1) open a pot lid, place seasoning into the pot, and close the lid; (2) turn off the stove, press the fire alarm, and grasp the fire extinguisher; (3) open a cabinet, retrieve an object, and close the cabinet; and (4) open a drawer, retrieve an object, and close the drawer. Each compositional trial is successful only when all constituent state transitions are completed in order.

The Real-world column of Table 2 reports the aggregate success rates for the action primitives and compositional tasks. These results summarize the complete perception-to-execution pipeline in a single benchmark table rather than repeating them in separate appendix tables.

I Details of Real-World Evaluation

The real-world experiments use a RealMan robot in a tabletop workspace containing the object set shown in Fig. 7. The smoke generator controls the smoke concentration during the real-world experiments, while an Intel RealSense D435 is mounted near the robot to provide the RGB observations used by the perception pipeline. The depth stream is not used at deployment. The same global-to-local observation procedure, SmokeScene and SmokeAfford models, primitive library, and closed-loop recovery policy are used for all real-world trials. The robot executes the selected 6-DoF

Table 6: Ablation study of the SmokeScene and SmokeAfford perception modules. Scene-understanding and affordance-prediction results are averaged across Low, Medium, and High smoke densities.

Ablation	**Metric 1**	**Metric 2**	**Metric 3**
Scene-understanding model (Avg over smoke densities): PQ↑/ PQ^{th}↑/ TR↑			
SmokeScene w/o local smoke severity weight	76.73	74.03	85.57
SmokeScene w/o encoder geometry supervision	76.31	73.58	84.87
SmokeScene w/o decoder geometry supervision	76.23	73.48	84.72
SmokeScene w/o local smoke severity weight + w/o decoder geometry supervision	75.95	73.15	84.01
SmokeScene (full)	**77.33**	**74.68**	**86.47**
Affordance-prediction model: KLD↓/ SIM↑/ NSS↑			
SmokeAfford w/o object mask refinement and encoding	2.6048	0.2166	6.4598
SmokeAfford w/o depth-guided feature modulation	0.9452	0.5262	14.7670
SmokeAfford w/o object-centric representation modeling	0.9709	0.5336	15.3084
SmokeAfford w/o object-conditioned affordance decoding	1.0231	0.5007	14.5104
SmokeAfford (full)	**0.9300**	**0.5420**	**16.0800**



Figure 6: **Real-world Evaluation**.

pose through the conditioned primitive trajectory, and success is determined from the object-state predicates described in Appendix H.

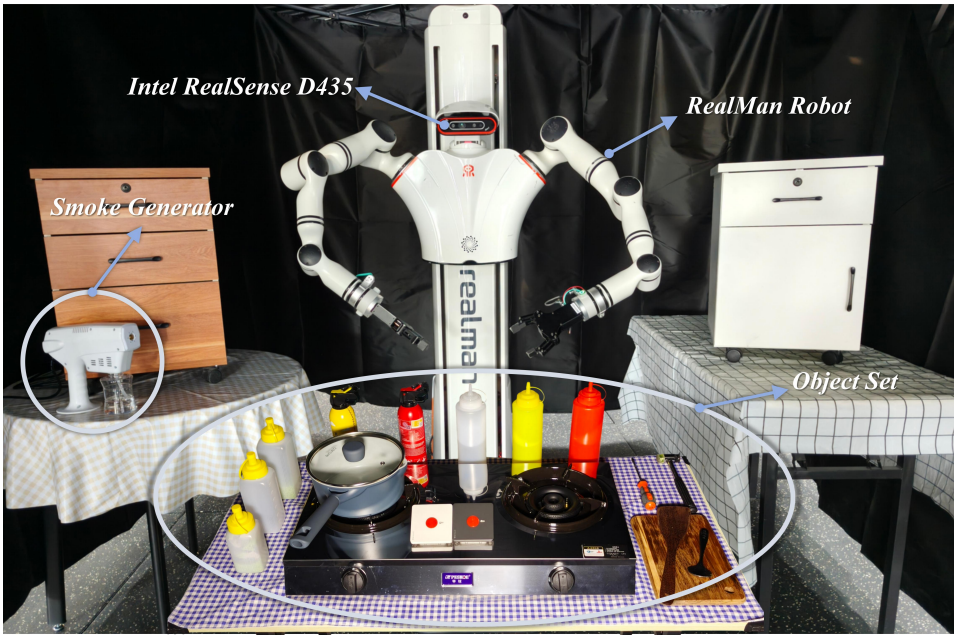


Figure 7: **Real-world evaluation setup**, including the RealMan robot, RGB camera, smoke generator for controlling experimental smoke concentration, and tabletop object set.