VLAff: Vision-Language-Affordance Model for Unified Actionable Affordances

Jihoon Oh[1], Kento Kawaharazuka[1], and Kei Okada[1]

***Abstract*— Learning manipulation skills from human videos is promising for scalable robot learning. However, the embodiment mismatch between humans and robots makes this challenging. One promising solution is to learn object-centric actionable affordances that are embodiment-agnostic. In this work, we propose a framework that leverages egocentric human videos with state-of-the-art 3D Structure-from-Motion and hand mesh reconstruction to extract actionable affordances such as visual, grasp, and trajectory affordances that explicitly encode where to interact, how to grasp, and how to move. We construct EgoAffordance, a large-scale dataset comprising 204K episodes with 5.6M visual affordances and 11.6M grasp and trajectory affordances. Building on this, we introduce VLAff, a large vision-language model-based unified foundation model that learns cross-modal correlations across all actionable affordances. Given a visual observation and instruction, VLAff generates visual affordance heatmaps, grasp poses, and trajectories, which are then converted into directly executable actions by utilizing 3D scene information. Through extensive experiments, we demonstrate that VLAff not only achieves state-of-the-art performance on visual affordance prediction, but can also be effectively applied to real robot applications such as zero-shot manipulation and affordance-guided robot learning.**

I. Introduction

Foundation models trained on large-scale datasets have proven effective for many downstream tasks in natural language processing and computer vision, such as LLaMA [12] and ChatGPT [13]. In robotics, there have been attempts to build such foundation models. However, traditional paradigms like imitation learning [14] face challenges: collecting robot action-state datasets is costly and time-consuming, with task and environment domains often limited to laboratory settings. In contrast, human video datasets [10], [8], [9], [11] can be easily collected from diverse environments and already exist at internet scale. However, direct use of human data is challenging due to embodiment differences between robots and humans, and video data lacks explicit action-state information. Previous work has attempted to learn implicit representations [15], [16], [17] from human videos for robot learning, but these are generally difficult to interpret and challenging to transfer directly into robot actions.

Affordances [18] offer a promising bridge between human demonstrations and robot task understanding. Many previous studies have attempted human-to-robot skill transfer using affordances such as heatmaps [19], [20], grasps [21], and trajectories [22] obtained from human demonstrations. However, these attempts rely on data acquired through sophisticated sensors in static laboratory environments. Recent works learn affordances from human videos, but typically focus on only one modality [5] or are limited to 2D image space [19]. There is no unified model that integrates visual, grasp, and trajectory affordances into a single framework.

In this work, we focus on (1) how to extract actionable affordance data from human videos, (2) how to train large-scale affordance knowledge in a single model, and (3) how to apply that affordance knowledge to downstream tasks like robot manipulation. We propose a pipeline that applies state-of-the-art 3D structure-from-motion to egocentric human videos to extract actionable affordance data using hand-object interaction detectors, segmentation models, and 3D hand reconstruction models, with inpainting to produce agent-agnostic scenes. As shown in Table I, our EgoAffordance dataset provides comprehensive actionable affordances compared to existing datasets. We then introduce a multi-modal learning framework based on large vision-language models that simultaneously predicts visual, grasp, and trajectory affordances from visual observations and instructions. Finally, we demonstrate effectiveness on downstream tasks such as zero-shot manipulation and affordance-guided learning.

II. Related Works

A. Learning from Human Videos

Human videos provide a scalable source for learning manipulation behaviors. Early approaches focused on learning visual representations through self-supervised learning [17], [16], [15], [23], [24]. These methods learn generalizable visual features from human videos that transfer to robotic tasks across diverse embodiments and environments. While these representation learning methods improve downstream task performance, they capture implicit visual patterns without explicit action information.

More recent work has shifted toward learning explicit action representations from human videos. Methods extract contact-based affordances [19], learn contact heatmaps and hand poses [25], or predict 3D hand trajectories [26] for robot pre-training and zero-shot manipulation. However, these methods focus on individual affordance modalities—visual contact points, hand poses, or trajectories—in isolation. Our work differs by jointly learning all three actionable affordances in a unified framework, enabling complete manipulation understanding from visual observations to grasp configurations and motion trajectories.

[1]The authors are with the University of Tokyo, Japan. {oh, kawaharazuka, k-okada}@jsk.imi.i.u-tokyo.ac.jp

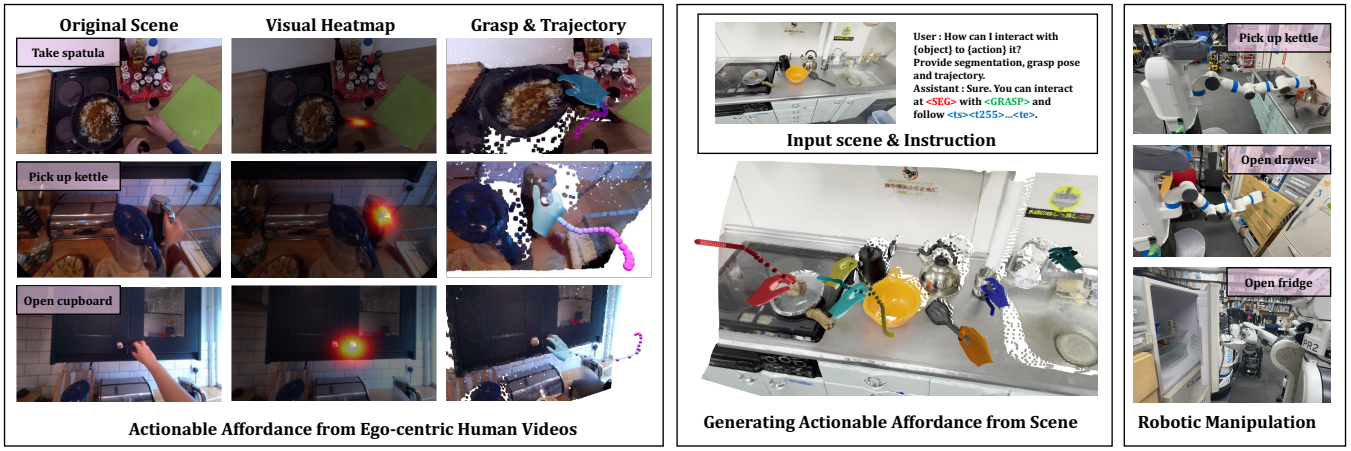

Fig. 1. We present **VLAff**, a Vision-Language-Affordance model that learns actionable affordances such as visual, grasp, trajectory from large-scale egocentric human videos to enable robot manipulation across diverse tasks.

TABLE I

COMPARISON OF AFFORDANCE-RELATED DATASETS. OUR EGOAFFORDANCE DATASET UNIQUELY PROVIDES COMPREHENSIVE ACTIONABLE AFFORDANCES INCLUDING VISUAL AFFORDANCE MASKS, HAND POSE ANNOTATIONS, AND CAMERA TRAJECTORIES, EXTRACTED FROM LARGE-SCALE EGOCENTRIC HUMAN VIDEOS. FRAME INDICATES THE TOTAL NUMBER OF FRAMES, INST (ACT) DENOTES THE NUMBER OF INSTANCES OR ACTIONS, AND OBJ REPRESENTS THE NUMBER OF OBJECT CATEGORIES.

Type	Dataset	Frame	Inst (Act)	Obj	Annotation		
					Contact	Hand Pose	Camera Traj
Visual Affordance	UMD [1]	30K	7	17	✓	✗	✗
	AGD20K [2]	23.8K	36	50	✓	✗	✗
	IIT-AFF [3]	8.8K	9	10	✓	✗	✗
	ADE-Aff [4]	10K	7	150	✓	✗	✗
	HOVA-500K [5]	500K	675	1.7K	✓	✗	✗
Egocentric HOI	H2O [6]	571K	36	8	✗	✓	✓
	HOI4D [7]	2.4M	800	16	✗	✓	✓
	EPIC-KITCHENS [8]	11.5M	125	331	✗	✗	✓
	HD-EPIC [9]	4.4M	1.2K	17K	✗	✗	✓
	Ego4D [10]	3.7M	1.7K	4.3K	✗	✗	✓
	Ego-Exo4D [11]	1.4M	4.5K	-	✗	✓	✓
Actionable Affordance	**EgoAffordance (Ours)**	**5.6M**	**1.7K**	**16.4K**	✓	✓	✓

B. Affordance Learning

a) Visual Affordance.: Visual affordances indicate where interactions occur on objects. Early works detected affordances from static images [1], [27], [28], [19] but were limited to simple object categories or closed-set affordances. Recent methods leverage large-scale datasets and achieve open-vocabulary affordance reasoning [20], [5], [29], [30], grounding affordances from natural language. However, these methods focus solely on visual interaction regions without modeling grasp configurations or motion trajectories.

b) Hand-Object Interaction.: Understanding hand-object interactions is crucial for manipulation learning. Datasets capture dexterous hand poses with RGB-D data [21], [31], detailed hand-object contact [32], large-scale interactions with affordance annotations [33], bimanual tool use [34], and 4D temporal modeling [7]. However, these datasets and methods focus on grasp pose and contact modeling in isolation, without integrating visual affordance prediction or trajectory generation for complete manipulation understanding.

Our work differs by jointly learning visual affordances, grasp configurations, and trajectories in a unified framework, capturing their cross-modal correlations for complete actionable affordance understanding.

C. Vision-Language-Action Models

Vision-language-action (VLA) models have transformed robotic manipulation through large-scale pretraining [35], [36], [37], [38], [39]. These models demonstrate that transformer architectures with vision-language pretraining achieve robust manipulation across multiple robot platforms and

embodiments. The development of large-scale robot datasets has been crucial for this progress. Open X-Embodiment [39] aggregates diverse robot manipulation data across multiple embodiments, while DROID [40] provides 76K in-the-wild manipulation trajectories across 564 scenes. These datasets enable training generalist policies that transfer across different robot platforms.

However, VLA models require robot-specific action labels and struggle with the embodiment gap from human demonstrations. Our work addresses this by learning actionable affordances as an intermediate representation, enabling effective transfer from human videos to robot control.

III. Affordance Extraction from Human Videos

A. Preliminaries

We formulate the task of learning actionable affordances from human videos as a multi-modal prediction problem. Given an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ and a natural language instruction $\mathbf{L}$, our goal is to predict three complementary object-centric affordance representations:

$$f_{\text{VLAff}}(\mathbf{I}, \mathbf{L}) = (\mathbf{A}_v, \mathbf{A}_g, \mathbf{A}_t) \tag{1}$$

where:

- $\mathbf{A}_v \in \mathbb{R}^{H \times W}$ is the visual affordance heatmap indicating the probability distribution of related regions in the scene.
- $\mathbf{A}_g \in \mathbb{R}^{96}$ is the grasp pose represented as MANO parameters [41], capturing the hand configuration during manipulation. This includes global wrist rotation and 15 local joint rotations, each represented using 6D rotation representation [42].
- $\mathbf{A}_t \in \mathbb{R}^{T \times 6}$ is the trajectory of the hand after interaction, where each waypoint includes both translation (3D) and rotation (3D), and T is the number of trajectory waypoints.

To generate object-centric actionable affordances, we first identify the peak interaction point from the visual affordance heatmap. We then project this 2D point to 3D space using the camera intrinsics and depth information to obtain 3D anchor point. The grasp pose and trajectory are subsequently transformed relative to this 3D anchor point through the spatial transformation, yielding the object-centric grasp pose and trajectory that align the predicted affordances to the object's coordinate frame centered at the interaction point. This object-centric formulation enables effective transfer of learned human manipulation strategies across different embodiments, as the complete affordance representation provides both the where (visual), how (grasp), and motion (trajectory) information necessary for execution.

B. Actionable Affordance Generation from Ego-Centric Video

a) Data Preparation: We begin by processing egocentric video datasets using their provided language descriptions and temporal annotations. For each video segment with language description $\mathbf{L}$, we sample frames $\{I_t\}$ around the annotated interaction timestamps. We query a pretrained Large VLM [13] on subsampled frames to extract key interaction information: contact keyframes $\{I_c\}$, action labels, object categories, and interacting hand information.

b) Hand-Object Interaction Extraction: For the identified contact keyframes $\{I_c\}$, we apply a hand-object detector [43] to obtain bounding boxes for hands and objects. We generate object masks [44] and extract 2D hand keypoints [45]. Contact regions are identified by computing the intersection between fingertip keypoints and object masks, then tracked [46].

For 3D hand reconstruction, we extract MANO parameters [47] representing hand pose and shape. To minimize visual bias from the ego perspective, we perform ego segmentation [48] followed by inpainting [49].

c) Ego-centric SfM: To reconstruct the 3D scene and trajectory information, we perform monocular depth estimation [50] on the inpainted frames. For videos lacking camera intrinsics, we estimate them [51]. We then estimate camera poses [52], filtering out dynamic regions using object and ego masks to ensure robust estimation.

With the estimated camera poses and depth maps, we track 3D hand trajectories [53], capturing both pre- and post-contact motion patterns essential for trajectory affordance learning.

IV. Vision-Language-Affordance Model

To enable unified learning of visual, grasp, and trajectory affordances, we extend a pre-trained vision-language model (VLM) with specialized affordance tokens. Our approach introduces three types of tokens to represent different affordance modalities:

Visual Affordance Token. We add a special token `<SEG>` to the VLM vocabulary to predict visual affordance heatmaps. Following the approach in LISA [54], the `<SEG>` token embedding is used to generate pixel-wise affordance probabilities through a segmentation decoder, producing the visual affordance heatmap $\mathbf{A}_v$.

Grasp Token. For grasp pose prediction, we introduce a `<GRASP>` token that encodes MANO hand parameters. The model learns to associate language instructions with corresponding hand configurations by fine-tuning the VLM to predict the 96-dimensional grasp affordance $\mathbf{A}_g \in \mathbb{R}^{96}$, which includes global wrist rotation and 15 local joint rotations represented using 6D rotation representation [42].

Trajectory Tokens. To handle trajectories autoregressively as discrete tokens, we employ spatial binning [35], [36], [37] to quantize the trajectory space. We discretize the 6D pose space (3D translation + 3D rotation) into discrete bins and create a vocabulary of trajectory tokens `<p0>`, ..., `<p`$_{N-1}$`>`, where each token corresponds to a specific bin in the quantized trajectory space. During training, continuous trajectory waypoints are converted to their nearest bin indices, and the model learns to predict sequences of trajectory tokens that can be decoded back to continuous 6D poses.

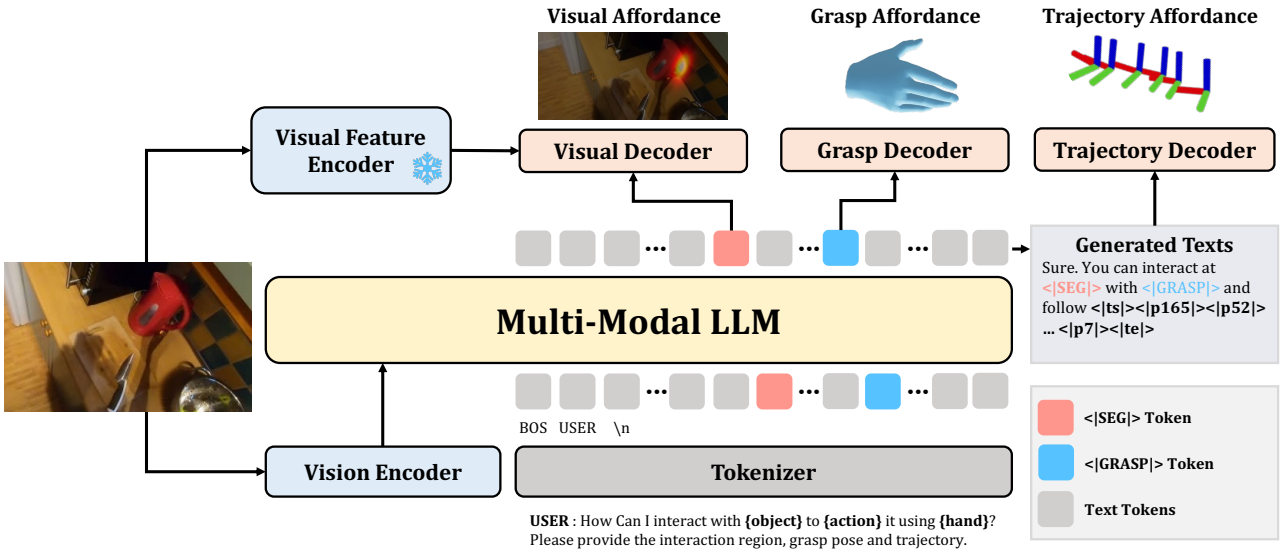

Fig. 2. VLAff model architecture showing the integration of vision encoder, VLM, and specialized decoders for visual, grasp, and trajectory affordance prediction.

By introducing these specialized affordance tokens, the model enables joint learning across all three affordance modalities and captures rich correlations between visual interaction regions, grasp configurations, and motion patterns.

A. Model Architecture

Our VLAff model consists of four main components that work together to predict comprehensive affordances from vision-language inputs. Figure 2 illustrates the overall architecture.

a) VLM: The core vision-language model processes the input image and language instruction to generate contextualized embeddings for all tokens in the sequence. The special affordance tokens <SEG>, <GRASP>, and <p⋆> are embedded within this sequence, producing corresponding token embeddings that are used by the specialized decoders.

For trajectory affordance prediction, the model generates trajectory tokens autoregressively. Given the context tokens up to position t, the model predicts the next trajectory token. This autoregressive generation continues until the model produces an end-of-trajectory token or reaches the maximum sequence length, resulting in a trajectory sequence that can be decoded back to continuous 6D poses.

b) Vision Encoder: While VLMs excel at general knowledge and reasoning capabilities, they often exhibit limitations in fine-grained visual grounding tasks that require precise spatial understanding. To address this limitation and enhance visual feature extraction beyond the standard VLM vision encoder, we incorporate an additional DINOv2 [55] vision encoder specifically designed for dense visual understanding. The DINOv2 encoder extracts rich visual features that provide fine-grained spatial information crucial for precise affordance localization, complementing the VLM's reasoning capabilities with stronger visual grounding abilities.

c) Visual Affordance Decoder: The visual affordance decoder fuses the enhanced visual features $\mathbf{F}_v$ from the DINOv2 encoder with the <SEG> token embedding $\mathbf{h}_{seg}$ to generate the visual affordance heatmap. Through a series of upsampling and fusion layers, the decoder produces the final visual affordance map $\mathbf{A}_v \in \mathbb{R}^{H \times W}$, where each pixel value represents the probability of being an interaction region.

d) Grasp Decoder: The grasp decoder takes the <GRASP> token embedding $\mathbf{h}_{grasp}$ and decodes it into MANO hand parameters. Specifically, it predicts the global wrist rotation and 15 local joint rotations using 6D rotation representation [42], forming the complete 96-dimensional grasp affordance $\mathbf{A}_g \in \mathbb{R}^{96}$.

B. Training Objectives

We train VLAff using specialized loss functions tailored to each affordance modality:

a) Visual Affordance Loss: For visual affordance prediction, we employ the Soft Dice Loss to handle the imbalanced nature of visual heatmap prediction tasks:

$$\mathcal{L}_{visual} = 1 - \frac{2\sum_{i,j}(\mathbf{A}_v^{pred}(i,j))^p \cdot (\mathbf{A}_v^{gt}(i,j))^p + s}{\sum_{i,j}(\mathbf{A}_v^{pred}(i,j))^p + \sum_{i,j}(\mathbf{A}_v^{gt}(i,j))^p + s} \tag{2}$$

where $\mathbf{A}_v^{pred}$ and $\mathbf{A}_v^{gt}$ are the predicted and ground truth visual affordance heatmaps, $s = 1.0$ is a smoothing factor, and $p = 1.5$ is a power parameter for soft boundary handling.

b) Grasp Affordance Loss: For grasp pose prediction, we use Smooth L1 loss:

$$\mathcal{L}_{grasp} = \frac{1}{J}\sum_{k=1}^{J} f(x_k), \quad x_k = \mathbf{A}_g^{pred}[k] - \mathbf{A}_g^{gt}[k] \tag{3}$$

where $J = 96$ and $f(x) = 0.5x^2$ if $|x| < 1$, else $|x| - 0.5$.

c) Trajectory Affordance Loss: For trajectory token prediction, we apply the standard Cross-Entropy loss used in autoregressive language models:

$$\mathcal{L}_{traj} = -\frac{1}{T}\sum_{t=1}^{T} \log P(\texttt{<pi>}_t^{gt} | \mathbf{I}, \mathbf{L}, \texttt{<p*>}_{1:t-1}) \quad (4)$$

where $\texttt{<pi>}_t^{gt}$ is the ground truth trajectory token at position t.

d) Total Loss: The total training loss combines all three objectives:

$$\mathcal{L}_{total} = \lambda_v \mathcal{L}_{visual} + \lambda_g \mathcal{L}_{grasp} + \lambda_t \mathcal{L}_{traj} \quad (5)$$

where λ_v, λ_g, and λ_t are weighting hyperparameters that balance the contribution of each affordance modality during training.

C. Trajectory Sampling Strategy

Our model infers trajectories from 2D images and language instructions, which poses challenges for generating physically plausible 3D trajectories. To address this limitation, we leverage the reasoning capabilities of our vision-language model backbone and introduce two strategies that guide the model to generate more plausible trajectories in 3D space during inference.

a) In-Context Trajectory Guidance: We utilize high-performance VLMs such as GPT [13] as a high-level planner to provide directional guidance for trajectory generation. Given the task prompt and interaction object, we query the VLM to provide a 3D directional vector $[d_x, d_y, d_z]$ in the ego-centric (camera) coordinate frame, where each component is constrained to $\{-1, 0, 1\}$ and encodes coarse motion intent along a camera axis: d_x horizontal (right $+1$), d_y vertical (down $+1$), and d_z depth (forward, away from camera, $+1$), with 0 meaning no motion on that axis. For example, $[0, 0, 1]$ indicates pushing forward, while $[1, 0, -1]$ indicates pulling backward and to the right. This guidance is applied only to the first trajectory token by sampling the initial xyz position tokens around the normalized and scaled directional vector:

$$\mathbf{p}_1 \sim \mathcal{N}(\mathbf{p}_{contact} + s \cdot \frac{\mathbf{d}_{guide}}{\|\mathbf{d}_{guide}\|}, \sigma^2 \mathbf{I}) \quad (6)$$

where $\mathbf{p}_{contact}$ is the interaction point, $\mathbf{d}_{guide}$ is the VLM-provided direction vector, s is a scale factor, and σ controls the sampling variance. This approach enables our model to produce trajectories that align with high-level task understanding while maintaining fine-grained spatial details learned from human demonstrations.

b) Sampling-Based Trajectory Selection: To generate physically plausible trajectories, we sample K trajectory candidates using nucleus sampling with top-p filtering, then select the optimal one using a composite objective that balances collision avoidance and directional alignment. For collision detection, we voxelize the 3D space and check trajectory waypoints against occupied voxels from the scene point cloud and depth map. The trajectory with the lowest combined objective is selected as the final output.

V. Experiments

We evaluate our VLAff model across four key aspects: (1) visual affordance prediction quality to assess how well the model localizes interaction regions, (2) trajectory generation effectiveness to validate our autoregressive generation strategy, (3) zero-shot manipulation performance to demonstrate direct transferability of generated affordances to robotic systems, and (4) affordance-guided policy learning to show how predicted affordances can serve as effective guidance for training environment-adaptive closed-loop policies.

A. Experimental Settings

a) Backbones: Our VLAff model leverages state-of-the-art foundation models as core components. We employ Qwen2.5-VL [56] as our base vision-language model. For enhanced visual feature extraction, we integrate DINOv2 [55] as our additional vision encoder.

b) Datasets: Our final EgoAffordance dataset comprises 204,025 episodes, containing 5,782,431 visual heatmaps and 11,612,524 trajectory sequences. To improve fine-grained visual affordance prediction, we additionally incorporate data from HANDAL [57] and SceneFun3D [58], which provide detailed object-level affordance annotations.

c) Visual Affordance: We evaluate visual affordance prediction performance on a test set consisting of 500 randomly sampled scenes from each egocentric video dataset. We compare VLAff with state-of-the-art segmentation-based methods: VRB [19], UAD [30], 3DOI [59], and LISA [54]. Note that LISA does not directly predict affordance heatmaps but is an LLM-based open vocabulary segmentation model; we normalize its output segmentation masks to [0,1] for fair comparison. We employ the following evaluation metrics: **IoU (Intersection over Union)** measures the overlap between predicted and ground truth affordance regions after normalizing heatmaps to [0,1] and applying a threshold of 0.5 to create binary masks; **NSS (Normalized Scanpath Saliency)** measures how well predicted affordances align with ground truth interaction points; **SIM (Similarity)** measures the similarity between predicted and ground truth affordance distributions; **KLD (Kullback-Leibler Divergence)** quantifies the divergence between predicted and ground truth affordance probability distributions.

d) Zero-Shot Manipulation: We evaluate VLAff's zero-shot manipulation capabilities against baseline methods on diverse manipulation tasks in both simulation and real-world environments. We compare against RAM [60], VRB [19], GeneralFlow [61], and VidBot [26]. For RAM, we sample retrieval episodes from our dataset. For VRB, we lift predicted 2D trajectories to 3D leveraging the strategy from RAM.

Since baseline methods do not directly output grasp poses, we employ GraspNet [62] to generate robot gripper pose for all baseline methods. For VLAff, we follow the retargeting strategy from previous work [63] to estimate gripper pose from predicted hand pose.

For simulation experiments, we use IsaacGym [64] as our simulation platform, with tasks from PartManip [65],

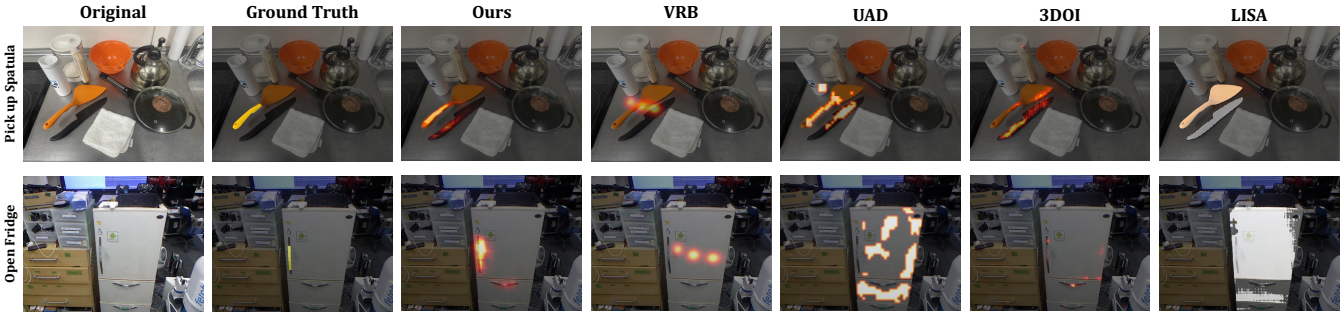

Fig. 3. Visual affordance prediction results in the wild. We show visual affordance predictions from VLAff and baseline methods on diverse manipulation scenarios. VLAff demonstrates superior localization of interaction regions compared to existing methods.

FrankaKitchen [66], and ManiSkill [67] that are included in the Ag2Manip [68] environment. We select 10 everyday household tasks that require understanding of object interaction spots and manipulation trajectories. These tasks encompass various action primitives such as opening, closing, pushing, and turning. Each method is tested 10 times per task without any task-specific training.

For real-world experiments, we deploy the generated affordances on mobile manipulation robots Fetch and PR2 to validate practical applicability. We evaluate on 5 manipulation tasks in real kitchen environments, testing each method 10 times per task.

e) Affordance-Guided Manipulation Learning: While VLAff demonstrates strong zero-shot manipulation capabilities, there exists a gap between affordances learned from videos and physical environments in real-world deployment. To bridge this gap, we employ real-world reinforcement learning guided by the predicted affordances, where learned actionable affordances provide structured priors to accelerate policy learning in the target robot embodiment. The reward function consists of two components: (1) a sparse task success reward determined by a vision-language model (VLM) that determines task success or failure, and (2) affordance-based guidance rewards that provide dense signals throughout the manipulation based on predicted contact points and trajectories (details in supplementary material). To evaluate the effectiveness of different affordance components as guidance for manipulation learning, we conduct an ablation study on the "Open Fridge" task in a real-world environment, comparing configurations using all affordance modalities versus ablated versions removing one modality at a time.

B. Experiment Results

1) Visual Affordance: Table II presents comprehensive visual affordance prediction results. VLAff achieves state-of-the-art performance across all segmentation-based metrics, demonstrating the effectiveness of our unified affordance learning approach. Beyond quantitative improvements, VLAff produces more fine-grained and natural affordance predictions compared to baseline methods, as illustrated in Figure 3. We analyze the limitations of baseline methods: VRB predicts discrete contact points rather than dense heatmaps, limiting fine-grained affordance localization; UAD, trained on rendered object images, struggles in cluttered real-world scenes; LISA, trained on object and instance-level segmentation data, tends to predict entire objects rather than specific contact points. We attribute VLAff's superior performance to two key factors: (1) the integration of DINOv2's part-aware visual features, which enable fine-grained localization of interaction regions, and (2) training on our large-scale, diverse EgoAffordance dataset spanning multiple domains and object categories. The results validate our key hypothesis that jointly learning visual, grasp, and trajectory affordances in a unified framework leads to better visual affordance prediction compared to methods that learn visual affordances in isolation.

TABLE II
VISUAL AFFORDANCE EVALUATION RESULTS

Method	IoU↑	NSS↑	SIM↑	KLD↓
VRB [19]	0.013	-0.032	0.052	6.942
UAD [30]	0.063	1.095	0.097	5.687
3DOI [59]	0.012	0.742	0.034	4.252
LISA [54]	0.113	1.342	0.025	3.886
Ours	**0.121**	**1.542**	**0.142**	**2.517**

2) Zero-Shot Manipulation: Table III presents the zero-shot manipulation results across diverse tasks in both simulation and real-world environments. In simulation tasks, VLAff achieves an average success rate of 83.0%, demonstrating strong generalization performance close to VidBot [26] (85.0%), the current state-of-the-art zero-shot manipulation framework. VidBot's strong overall performance in simulation can be attributed to its ability to access spatial input such as depth, enabling the model to directly leverage spatial characteristics for manipulation planning, which is particularly beneficial in controlled simulation environments.

Notably, VLAff demonstrates superior performance on real-world tasks, achieving an average success rate of 68.0%, 16 percentage points higher than VidBot's 52.0%. This performance gap validates our key design choice of training on large-scale real-world egocentric video datasets, which better capture the complexity and diversity of real manipulation scenarios, while other methods frequently fail to estimate appropriate interaction regions in cluttered real-world scenes. VLAff's unified affordance learning approach enables more

TABLE III

ZERO-SHOT MANIPULATION PERFORMANCE. SUCCESS RATES AVERAGED OVER 10 TRIALS PER TASK. **SIM (T01-T10): T01**: OPEN HINGE CABINET, **T02**: OPEN SLIDE CABINET, **T03**: OPEN MICROWAVE, **T04**: CLOSE MICROWAVE, **T05**: PICK UP KETTLE, **T06**: OPEN DISH WASHER, **T07**: LIFT LID, **T08**: PULL DRAWER, **T09**: CLOSE HINGE CABINET, **T10**: TURN FAUCET. **REAL (T11-T15): T11**: OPEN DRAWER, **T12**: PICK UP BUCKET, **T13**: OPEN POT LID, **T14**: TAKE PAN, **T15**: PICK UP KETTLE.

Method	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	Avg	T11	T12	T13	T14	T15	Avg
RAM [60]	40	20	50	80	90	40	40	60	90	0	51.0	60	20	20	20	0	24.0
VRB [19]	20	10	60	70	90	30	50	50	80	20	48.0	60	40	20	20	0	28.0
GFlow [61]	10	10	60	**100**	90	60	70	50	90	10	55.0	80	20	20	20	0	28.0
VidBot [26]	**70**	**80**	**80**	**100**	**100**	70	**100**	**100**	**100**	50	**85.0**	**100**	**60**	**60**	40	0	52.0
Ours	**70**	60	70	**100**	**100**	**80**	90	**100**	**100**	**60**	83.0	**100**	**60**	**80**	**60**	**40**	**68.0**

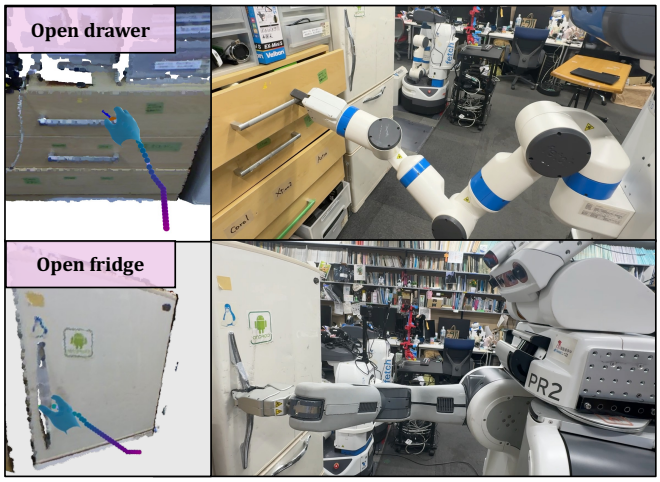

Fig. 4. Robot manipulation experiments. We visualize the generated affordances and their application to real robot manipulation tasks.

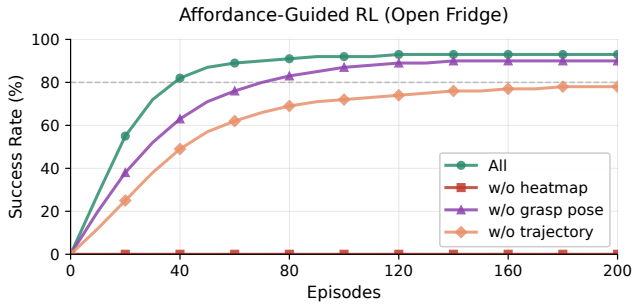

Fig. 5. Affordance-guided manipulation learning results on real-world Open Fridge task. We compare learning efficiency when using different combinations of affordance guidance

robust contact interaction prediction through joint learning of visual, grasp, and trajectory affordances. Our failure cases primarily stem from incorrect initial direction sampling during trajectory generation, which can be improved through better integration of scene context and spatial reasoning.

3) Affordance-Guided Manipulation Learning: Figure 5 shows the learning curves for each configuration. The complete model with all affordance modalities achieves the best performance most efficiently, validating that visual heatmap, grasp pose, and trajectory information provide complementary guidance. Most critically, when the visual affordance heatmap is removed, the model completely fails to achieve successful manipulation regardless of the number of training episodes, demonstrating that visual affordance is absolutely essential for manipulation learning as it provides the most important cues for object interaction by indicating where to make contact. Interestingly, grasp pose plays a less significant role in manipulation learning, which we attribute to the fundamental morphological differences between human hands and robot grippers—the exact hand configuration from human demonstrations does not directly transfer to gripper-based manipulation.

VI. CONCLUSION

We built EgoAffordance, a large-scale actionable affordance dataset comprising visual heatmaps, grasp poses and manipulation trajectories extracted from egocentric human videos leveraging state-of-the-art computer vision techniques. Upon this dataset, we presented VLAff, a novel Vision-Language-Affordance model that, unlike previous methods treating visual, grasp, and trajectory affordances separately, provides comprehensive action guidance by jointly learning these modalities within a unified VLM framework to predict actionable affordances from language instructions. We demonstrate through extensive experiments on visual affordance prediction, zero-shot manipulation, and affordance-guided manipulation learning that VLAff achieves results comparable to or surpassing existing state-of-the-art methods, validating the effectiveness of our unified affordance learning approach for robotic manipulation tasks. While VLAff demonstrates strong performance, a limitation remains: occasional generation of implausible trajectories that may not respect physical constraints. Future work could address this through incorporation of 3D scene-aware model architectures [69], as demonstrated in recent work [70].

REFERENCES

[1] A. Myers *et al.*, "Affordance detection of tool parts from geometric features," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1374–1381.

[2] H. Luo *et al.*, "Learning affordance grounding from exocentric images," in *CVPR*, 2022, pp. 2252–2261.

[3] A. Nguyen *et al.*, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 5908–5915.

[4] C.-Y. Chuang *et al.*, "Learning to act properly: Predicting and explaining affordances from images," in *CVPR*, 2018, pp. 975–983.

[5] T. Ma *et al.*, "Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation," *arXiv preprint arXiv:2505.11865*, 2025.

[6] T. Kwon *et al.*, "H2o: Two hands manipulating objects for first person interaction recognition," in *ICCV*, 2021, pp. 10 138–10 148.

[7] Y. Liu *et al.*, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *CVPR*, 2022, pp. 21 013–21 022.

[8] D. Damen *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018, pp. 720–736.

[9] H. Doughty *et al.*, "Hd-epic: A highly-detailed egocentric video dataset," *arXiv preprint arXiv:2502.04144*, 2025.

[10] K. Grauman *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *CVPR*, 2022, pp. 18 995–19 012.

[11] K. Grauman *et al.*, "Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives," *arXiv preprint arXiv:2311.18259*, 2023.

[12] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[13] OpenAI *et al.*, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[14] D. A. Pomerleau, "Efficient training of artificial neural networks for autonomous navigation," *Neural Computation*, vol. 3, no. 1, pp. 88–97, 1991.

[15] Y. J. Ma *et al.*, "Vip: Towards universal visual reward and representation via value-implicit pre-training," *arXiv preprint arXiv:2210.00030*, 2023.

[16] T. Xiao *et al.*, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.

[17] S. Nair *et al.*, "R3m: A universal visual representation for robot manipulation," *arXiv preprint arXiv:2203.12601*, 2022.

[18] J. J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

[19] S. Bahl *et al.*, "Affordances from human videos as a versatile representation for robotics," in *CVPR*, 2023, pp. 13 778–13 790.

[20] T. Ma *et al.*, "Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping," *arXiv preprint arXiv:2411.12286*, 2024.

[21] Y.-W. Chao *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *CVPR*, 2021, pp. 9044–9053.

[22] K. Shaw *et al.*, "Videodex: Learning dexterity from internet videos," in *Conference on Robot Learning*, 2023, pp. 654–665.

[23] A. Majumdar *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *arXiv preprint arXiv:2303.18240*, 2023.

[24] Y. J. Ma *et al.*, "Liv: Language-image representations and rewards for robotic control," *arXiv preprint arXiv:2306.00958*, 2023.

[25] M. K. Srirama *et al.*, "Hrp: Human affordances for robotic pre-training," in *Proceedings of Robotics: Science and Systems (RSS)*, Delft, Netherlands, 2024.

[26] Z. Han *et al.*, "Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation," *arXiv preprint arXiv:2503.07135*, 2025.

[27] J. Sawatzky *et al.*, "Weakly supervised affordance detection," in *CVPR*, 2017, pp. 2795–2804.

[28] G. Li *et al.*, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," in *CVPR*, 2023, pp. 10 922–10 931.

[29] S. Qian *et al.*, "Affordancellm: Grounding affordance from vision language models," in *CVPR*, 2024, pp. 7587–7597.

[30] Y. Tang *et al.*, "Uad: Unsupervised affordance distillation for generalization in robotic manipulation," *arXiv preprint arXiv:2506.09284*, 2025.

[31] Z. Fan *et al.*, "Arctic: A dataset for dexterous bimanual hand-object manipulation," in *CVPR*, 2023, pp. 927–938.

[32] S. Brahmbhatt *et al.*, "Contactpose: A dataset of grasps with object contact and hand pose," in *ECCV*, 2020.

[33] L. Yang *et al.*, "Oakink: A large-scale knowledge repository for understanding hand-object interaction," in *CVPR*, 2022.

[34] Y. Li *et al.*, "Taco: Benchmarking generalizable bimanual tool-action-object understanding," *arXiv preprint arXiv:2401.08399*, 2024.

[35] A. Brohan *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[36] A. Brohan *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[37] M. J. Kim *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[38] O. M. Team *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[39] A. Padalkar *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[40] A. Khazatsky *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," in *Robotics: Science and Systems*, 2024.

[41] J. Romero *et al.*, "Embodied hands: Modeling and capturing hands and bodies together," in *ACM TOG*, vol. 36, no. 6, 2017, pp. 1–17.

[42] Y. Zhou *et al.*, "On the continuity of rotation representations in neural networks," in *CVPR*, 2019, pp. 5745–5753.

[43] D. Shan *et al.*, "Understanding human hands in contact at internet scale," in *CVPR*, 2020, pp. 9869–9878.

[44] N. Ravi *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[45] Y. Xu *et al.*, "Vitpose: Simple vision transformer baselines for human pose estimation," in *NeurIPS*, 2022, pp. 38 571–38 584.

[46] N. Karaev *et al.*, "Cotracker3: Simpler and better point tracking by pseudo-labelling real videos," *arXiv preprint arXiv:2410.11831*, 2024.

[47] R. A. Li and D. Shan, "Wilor: End-to-end 3d hand localization and reconstruction in-the-wild," *arXiv preprint arXiv:2407.10034*, 2024.

[48] B. Zhang *et al.*, "Egohos: Dataset and method for hand and object segmentation in egocentric videos," in *CVPR*, 2022, pp. 21 381–21 391.

[49] S. Zhou *et al.*, "Propainter: Improving propagation and transformer for video inpainting," *arXiv preprint arXiv:2309.03897*, 2023.

[50] R. Wang *et al.*, "Moge-2: Accurate monocular geometry with metric scale and sharp details," *arXiv preprint arXiv:2507.02546*, 2025.

[51] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[52] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," in *NeurIPS*, 2021, pp. 16 558–16 569.

[53] A. W. Harley *et al.*, "Tapip3d: Tracking any point in persistent 3d geometry," *arXiv preprint arXiv:2312.03904*, 2024.

[54] X. Lai *et al.*, "Lisa: Reasoning segmentation via large language model," in *CVPR*, 2024, pp. 9579–9589.

[55] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[56] S. Bai *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[57] A. Guo *et al.*, "Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 428–11 435.

[58] A. Delitzas *et al.*, "Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes," in *CVPR*, 2024.

[59] S. Qian and D. F. Fouhey, "Understanding 3d object interaction from a single image," in *ICCV*, 2023, pp. 21 753–21 763.

[60] Y. Kuang *et al.*, "Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation," *arXiv preprint arXiv:2407.04689*, 2024.

[61] Z. Yuan *et al.*, "Generalflow: Generalizable manipulation policy with flow matching," *arXiv preprint arXiv:2410.10649*, 2024.

[62] H.-S. Fang *et al.*, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020.

[63] G. Papagiannis *et al.*, "R+x: Retrieval and execution from everyday human videos," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[64] V. Makoviychuk *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[65] H. Geng *et al.*, "Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations," in *CVPR*, 2023.

[66] A. Gupta *et al.*, "Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning," in *Conference on Robot Learning (CoRL)*, 2019.

[67] J. Gu *et al.*, "Maniskill2: A unified benchmark for generalizable manipulation skills," in *ICLR*, 2023.

[68] H. Geng *et al.*, "Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[69] R. Xu *et al.*, "Pointllm: Empowering large language models to understand point clouds," in *ECCV*, 2024.

[70] T. Yoshida *et al.*, "Generating 6dof object manipulation trajectories from action description in egocentric vision," in *CVPR*, 2025.