Foundation ViTs Can Learn Where to Act: Task Affordance Segmentation and Surface Normals

Francesco Dorati[1⋆] and Tommaso Felice Banfi[2]

[1] Politecnico di Milano, Milan, Italy
francesco.dorati@mail.polimi.it
[2] Talos Robotics AI, Milan, Italy
tommasofelice.banfi@talosrobotics.ai

Abstract. A robot given a task description must identify not only *what* to do, but also *where* the interaction is possible and *how* to approach the object. We formulate this problem through task affordances, represented as functional regions coupled with their local surface geometry. We show that a frozen foundation vision model can provide this grounding signal through a lightweight task-specific readout: a DINOv2 ViT-B/14 backbone with a 5.6M-parameter decoder trained from scratch (6.5% of the backbone size). On the UMD Part-Affordance benchmark, our method achieves a Jaccard score of 80.29 under the official evaluation protocol, improving over the previous best architecture by 4.4 points (82.92 without the normal prediction branch), while requiring neither object detectors, bounding boxes, nor object identity. Beyond affordance segmentation, the same model predicts a dense surface-normal field with a mean angular error of 21.6°, providing geometric cues that could support interaction planning, without additional normal annotations. We further study the role of affordance representations in vision-language-action models: probing the humanoid VLA Ψ_0 on 340 goal-level commands, a linear classifier recovers the target affordance with 84.6% accuracy at mid layers, while visual input adds up to 10.2 points over text alone. These results suggest that ViTs can serve as perceptual interfaces for VLA systems by grounding task descriptions into action-relevant areas.

Keywords: Affordance Segmentation · Vision Transformer · Robotics · Vision-Language-Action Models · Surface Normals · Computer Vision

1 Introduction

An object category label does not tell a robot what to do. A knife invites two incompatible interactions: grasping its handle, cutting with its blade. Both those actions need the functional region's extent and the local surface orientation to approach it. This is Gibson's *affordance* [13], the minimal interface between perception and planning; we use *grasp packet* to denote a perceptual output: a task-selected region paired with a local approach-direction cue (Sec. 3). This new representation also provides a target for future synthetic VLA pre/post-training.

⋆ Corresponding author: F. Dorati (francesco.dorati@mail.polimi.it)

A decade of specialized architectures [3,10] has advanced affordance segmentation, yet under a fair evaluation protocol the strongest method reaches only 75.86 Jaccard on UMD [2,3,23]. Rather than a new end-to-end network, we train a *compact readout* (5.6M parameters) over a frozen self-supervised DINOv2 backbone [27], spending capacity only where the frozen prior is weak (token fusion, a spatial-detail encoder, an imbalance-aware loss, equivariant supervision). This attains the best result under the benchmark's own scorer, and the *same* forward pass also emits the approach geometry.

We argue this is the perceptual interface that vision-language-action models (VLAs) [6, 35] need both at training and inference time. A fast-growing line of work [18, 24, 31, 34] already treats affordance as the intermediate representation between goal-level language and low-level control, but *trains* it into the policy while still consuming a where-to-act signal some module must produce. Our grasp packet is produced by a light decoder on a frozen backbone, with no policy retraining or robot data. A downstream policy could use the selected affordance map and local normal as contact-region and orientation cues. We do not evaluate grasp planning, execution, collision avoidance, or safety here. This presumes the coupling is well posed and that a VLA *already* encodes the affordance even when language never names it ("prepare the tomato" $\implies$ `cut`). Probing VLA Ψ_0 [32] against its base VLM settles it: action pretraining grounds the affordance in the image (+10 points); vision-language pretraining alone does not (+0.5).

Contributions:

1. **A trained multi-task affordance network on UMD.** A 5.6M-parameter decoder over frozen DINOv2 reaches 80.29 Jaccard (Sec. 6) on the held-out test split under the benchmark's released scorer (+4.4 over M2F-AFF [3]).
2. **The grasp packet: segmentation and local geometry from one forward pass.** From the *same* forward pass, we predict a dense surface-normal field that can serve as a local orientation cue. Its 21.6° mean error is within a few degrees of a 15.5° depth-supervision consistency estimate; VLA manipulation improvement is not evaluated here, but is planned as future work.
3. **A probing method and findings for VLA affordance grounding.** With matched text-/image-only controls affordances are *linearly* decodable from frozen Ψ_0 embeddings (up to 84.6%), and visual input adds up to 10.2% over the instruction-only condition. This affordance awareness contribution emerges from action post-training, not vision-language pretraining (Sec. 7.1).

2 Related Work

2.1 Affordance Segmentation

The UMD Part Affordance dataset [23] remains the reference benchmark for part-level affordance segmentation: ∼30K RGB-D turntable images of 105 tool instances in 17 categories, pixel-annotated with seven affordance classes (`grasp`,

`cut`, `scoop`, `contain`, `pound`, `support`, `wrap-grasp`); the official category split holds out roughly half the instances per category (14,823 train / 14,020 test), so test tools are never seen during training. Myers et al.'s original detectors used hand-crafted geometric features over superpixels [23]; the first deep models replaced them with encoder–decoder CNNs [26]. AffordanceNet [10] introduced the dominant two-stage design, a Faster-R-CNN-style detector [29] with an RoI-aligned affordance mask branch, which requires object-level boxes and identities at training time. DRNAtt [14] kept an encoder–decoder but added dilated residual features with spatial and channel attention, and M2F-AFF [3] adapted the query-based universal segmenter Mask2Former [8] to affordance classes. A parallel weakly-supervised line grounds affordances without pixel masks: LOCATE [19] transfers interaction cues from exocentric demonstration images to egocentric test images using only image- and object-level labels.

Two properties distinguish our model from the fully-supervised detectors above: it is *multi-label* rather than single-label, since overlapping affordances are the physical reality (Sec. 3), and it trains *no backbone*, only a compact decoder (Sec. 4), with no detection stage, box supervision, or object identity.

The reproducibility problem. Apicella et al. [2, 3] re-implemented and re-benchmarked the field under one protocol (same split, resolution handling and metric code), releasing per-image outputs and the `aff-seg-eval` scorer. Under this protocol the UMD ranking (Tab. 2) sits far below some originally published figures (*e.g.* AffordanceNet's $F_\beta^\omega = 0.799$, obtained at different resolution with a detection pipeline). We adopt their toolkit for results comparison (Sec. 5).

2.2 Frozen Foundation Features for Dense Prediction

DINOv2 [27] distills a self-supervised ViT from 142M curated images and reports strong frozen-feature transfer to segmentation and depth with linear or light heads; register tokens [9] further clean its dense feature maps. DPT [28] established multi-depth token fusion for dense prediction, and skip connections from the input resolution date back to U-Net [30]. Our decoder combines both ideas under a constraint none of these works impose: the *entire* semantic encoder stays frozen, with all spatial detail re-injected by a spatial-detail encoder trained from scratch (Sec. 4.1). We adopted promptable segmenters (SAM [17]) and open-set detectors (Grounding DINO [20]) only for ablation analysis (Sec. 6.3).

2.3 Affordance as the Intermediate Representation for VLAs

VLAs map (image, language instruction) pairs to low-level actions [6,35]; Ψ_0 [32] targets humanoid loco-manipulation by autoregressively pretraining a Qwen3-VL-2B backbone [5] on large-scale egocentric human video before attaching a flow-based action expert. Because raw VLM features emphasize global object appearance rather than the functional region a task acts on [18], a rapidly growing line inserts *affordance as an explicit intermediate representation* between

language and control. RT-Affordance [24] lifts novel-task success by over 50% using affordances as lightweight, transferable plans; AffordanceVLA [34] forecasts a where/how affordance inside a retrained policy; Afford-VLA [31] conditions action generation on decoded affordance tokens; AffordVLA [18] distills a zero-shot affordance teacher into the policy's visual features; and RAG-Net [33]/ManipGPT [16] likewise couple affordance segmentation to grasping; GLOVER++ [21] distills affordance knowledge from large-scale human demonstration video into open-vocabulary manipulation reasoning. All *train* affordance into the policy (via mixture-of-transformer retraining, joint action–affordance optimization, teacher distillation, or robot-trajectory supervision), yet still consume a where-to-act signal some upstream module must produce; our frozen-backbone decoder could be that module. All also assume the affordance is *absent* from the policy; we test the converse: does a frozen, already-trained VLA already partially encode the task-relevant affordance, and which pretraining stage put it there? Linear probing is the standard tool [1], but to our knowledge has not been applied to VLA affordance analysis with matched text- and image-only controls.

3 Problem Formulation

3.1 Grasp-Packet Perception

Let $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{Y}_i, Z_i)\}$ pair RGB images $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ with affordance annotations $\mathbf{Y}$ and depth Z (supervision only; absent at test). The C=7 classes are *not* mutually exclusive: a mug exposes `wrap-grasp` (handle), `contain` (cavity), and `grasp` (rim) at once, so the target is a multi-label tensor $\mathbf{Y} \in \{0,1\}^{C\times H\times W}$ and any single-label softmax misspecifies the task. The model is a map

$$f_\theta : \mathbf{I} \longmapsto \big(\hat{\mathbf{M}}, \hat{\mathbf{N}}\big), \qquad \hat{\mathbf{M}} \in [0,1]^{C\times H\times W}, \quad \hat{\mathbf{N}}(\mathbf{x}) \in \mathbb{S}^2, \tag{1}$$

producing per-class occupancy and a unit normal field; only the readout θ is learned, the encoder is fixed (Sec. 4). At deployment, given the task class c, intrinsics K, and a depth reading, the dense outputs collapse to the *grasp packet*

$$g_c = \big(\mathbf{p}_c, \mathbf{n}_c\big) \in \mathbb{R}^3 \times \mathbb{S}^2, \qquad \mathbf{p}_c = Z(\mathbf{u}_c)\, K^{-1}\tilde{\mathbf{u}}_c, \quad \mathbf{n}_c = \hat{\mathbf{N}}(\mathbf{u}_c), \tag{2}$$

with $\mathbf{u}_c$ the centroid of $\{\hat{M}_c > \tau\}$.

The UMD protocol scores single-label maps (Sec. 5), so for evaluation only we have to collapse predictions by per-pixel arg-max above a background threshold

$$\hat{y}(\mathbf{x}) = \begin{cases} \arg\max_c \hat{M}_c(\mathbf{x}) & \max_c \hat{M}_c(\mathbf{x}) > \tau \\ 0 \text{ (background)} & \text{otherwise,} \end{cases} \qquad \tau = 0.5, \tag{3}$$

3.2 Affordance Grounding as a Decoding Problem

Let f be a VLA whose multimodal backbone produces hidden states $\mathbf{e}^{(\ell)}(\mathbf{I}, T) \in \mathbb{R}^d$ at layers $\ell = 1, \ldots, L$ for an image and a *goal-level* instruction T (one that

expresses an intent without naming the affordance or its verb), annotated with the affordance $a^*(\mathbf{I}, T) \in \mathcal{A}$ the task requires. We measure the *decodability* of a^* from the frozen representation through a probe

$$g_\phi(\mathbf{e}) = \mathrm{softmax}(W\mathbf{e} + b), \qquad \hat{a} = \arg\max_{a \in \mathcal{A}} g_\phi(\mathbf{e})_a, \tag{4}$$

trained with cross-entropy on $\{(\mathbf{e}_i^{(\ell)}, a_i^*)\}$ and evaluated on instruction *phrasings* disjoint from training. Three input conditions, `full` $(\mathbf{I}, T)$, `instruction_only` (T), and `image_only` $(\mathbf{I})$, decompose the sources of decodability, and the *visual contribution* at layer ℓ,

$$\Delta^{(\ell)} = \mathrm{Acc}^{(\ell)}_{\texttt{full}} - \mathrm{Acc}^{(\ell)}_{\texttt{instruction\_only}}, \tag{5}$$

isolates what the image adds beyond the instruction's lexical content at the same layer and split. Two properties make the protocol identifying: the split must hold out *phrasings* (holding out object instances leaks identical strings and drives $\mathrm{Acc}_{\texttt{instruction\_only}}$ to 96.5%, erasing Δ), and the control model must share the backbone: we contrast Ψ_0 against the base Qwen3-VL-2B it was pretrained from, so any difference in Δ is attributable to action pretraining alone.

4 Method

4.1 Architecture

We use a self-supervised transformer as a *fixed visual prior* and cast affordance perception as learning a lightweight *decoding operator* onto it. Formally we factor the predictor as $f_\theta = R_\theta \circ \Phi$, so Φ is the frozen backbone and every trained parameter lives in R_θ. A DINOv2 ViT-B/14 with registers Φ (86M parameters, permanently frozen) maps $\mathbf{I} \in \mathbb{R}^{3\times448\times448}$ to patch-token grids at four of its twelve blocks (see Figure 1),

$$\mathbf{Z}_\ell = \Phi_\ell(\mathbf{I}) \in \mathbb{R}^{768\times32\times32}, \qquad \ell \in \{2, 5, 8, 11\}, \tag{6}$$

capturing local structure at early depths before self-attention diffuses it and semantics at late layers [28]. Two deficiencies remain. (i) *Spatial bandwidth*: a 32×32 grid cannot encode 448×448 boundaries, and no decoder can recover information the encoding discarded. We therefore train a four-stage spatial-detail encoder σ from scratch on the raw image, $\sigma(\mathbf{I}) = (\mathbf{S}_0, \ldots, \mathbf{S}_3)$ with $\mathbf{S}_j \in \mathbb{R}^{c_j\times448/2^j\times448/2^j}$, $c = (32, 64, 96, 128)$, whose only job is to carry edges and texture to the decoder. (ii) *Task alignment*: token channels are generic. Four learned 1×1 projections ρ_ℓ and a fusion block ψ recombine them,

$$\mathbf{X}_0 = \psi\big([\rho_2\mathbf{Z}_2; \rho_5\mathbf{Z}_5; \rho_8\mathbf{Z}_8; \rho_{11}\mathbf{Z}_{11}]\big) \in \mathbb{R}^{256\times32\times32}, \tag{7}$$

after which four fusion stages interleave parameter-free bilinear upsampling $\mathcal{U}$ with skip concatenation and learned mixing,

$$\mathbf{X}_k = g_k\big([\mathcal{U}(\mathbf{X}_{k-1}); \mathbf{S}_{4-k}]\big), \qquad k = 1, \ldots, 4, \tag{8}$$

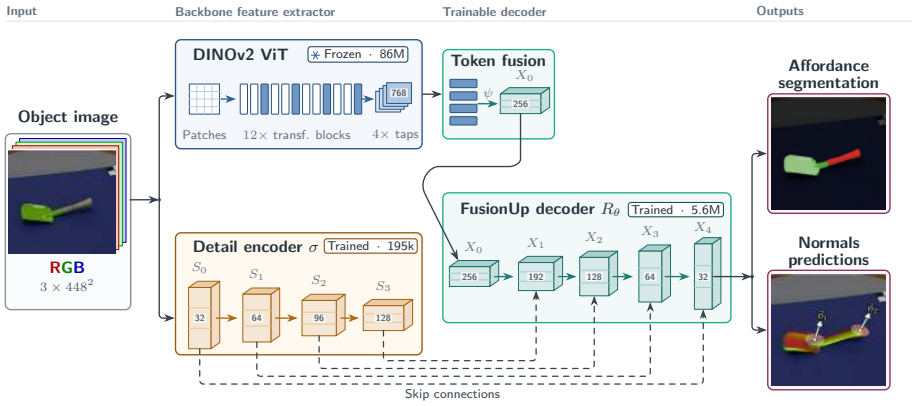

Fig. 1: Architecture. A frozen DINOv2 ViT-B/14 supplies patch tokens from four depths (layers 2/5/8/11); a from-scratch spatial-detail encoder supplies high-frequency skip features at four resolutions; the trained fusion decoder (5.6M parameters total) outputs affordance segmentation and a surface-normal field at 448×448.

(ψ, g_k: double conv–BN–ReLU; bilinear $\mathcal{U}$ avoids transposed-convolution checkerboards). The shared trunk $\mathbf{X}_4 \in \mathbb{R}^{32\times448\times448}$ feeds two heads:

$$\mathbf{A} = W_a * \mathbf{X}_4 \in \mathbb{R}^{7\times448\times448}, \qquad \hat{\mathbf{M}} = \sigma_{\text{sig}}(\mathbf{A}), \tag{9}$$

$$\hat{\mathbf{N}} = \tilde{\mathbf{N}}/\|\tilde{\mathbf{N}}\|_2, \qquad \tilde{\mathbf{N}} = W_{n_2} * \phi\big(W_{n_1} * \mathbf{X}_4\big), \tag{10}$$

with independent sigmoids σ_{sig} per affordance channel (the multi-label structure of Sec. 3) and per-pixel ℓ_2 normalization projecting normals onto $\mathbb{S}^2$. The trained parameters $\theta = \{\sigma, \rho_{2..11}, \psi, g_{1..4}, W_a, W_{n_{1,2}}\}$ are 5,6M (6.5% of the backbone).

4.2 Training Objective

Every affordance class occupies <1% of training pixels (from `contain` 0.67% down to `scoop` 0.14%), so the objective must defeat both the all-background solution and gradient domination by frequent classes. We optimize

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{seg}} + \lambda_n \mathcal{L}_{\text{normal}} + \lambda_s \mathcal{L}_{\text{smooth}}, \qquad \lambda_n = 5,\ \lambda_s = 0.5. \tag{11}$$

Segmentation Loss $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{wBCE}} + \mathcal{L}_{\text{Dice}}$ combines logit-space cross-entropy, reweighted per class by an inverse-frequency factor computed once from training pixel counts $N_c^{\pm}$ and $w_c = \min\Big(\sqrt{N_c^- / N_c^+},\, 15\Big)$, so the loss $\mathcal{L}_{\text{wBCE}}$ is defined as

$$\mathcal{L}_{\text{wBCE}} = -\frac{1}{CHW} \sum_{c,\mathbf{x}} \Big[w_c\, Y_c \log \hat{M}_c + (1 - Y_c) \log(1 - \hat{M}_c) \Big], \tag{12}$$

with a soft Dice term computed *per channel* on the spatial dimensions and averaged so that rare classes carry equal weight,

$$\mathcal{L}_{\text{Dice}} = \frac{1}{C}\sum_{c}\left[1 - \frac{2\langle \hat{\mathbf{M}}_c, \mathbf{Y}_c\rangle + 1}{\|\hat{\mathbf{M}}_c\|_1 + \|\mathbf{Y}_c\|_1 + 1}\right]. \tag{13}$$

The square root in Eq. (12) tempers raw ratios exceeding 10^4 and the clip bounds gradient scale; five of seven classes saturate it ($w_{\texttt{grasp}} = 12.75$, $w_{\texttt{contain}} = 12.33$, etc) helping the rare ones avoiding train–validation gaps (`support` and `scoop`).

Normals Loss Normals are supervised by cosine dissimilarity over the annotated region $\Omega = \{\mathbf{x} : \sum_c Y_c(\mathbf{x}) > 0\}$, the surfaces a robot will contact, with an edge-aware prior regularizing the field elsewhere:

$$\mathcal{L}_{\text{normal}} = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}\big(1 - \langle \hat{\mathbf{N}}, \mathbf{N}^*\rangle\big), \qquad \mathcal{L}_{\text{smooth}} = \sum_{d\in\{u,v\}}\big\langle |\partial_d \hat{\mathbf{N}}|\, e^{-\gamma|\partial_d \mathbf{I}|}\big\rangle, \quad \gamma = 10. \tag{14}$$

4.3 Surface-Normal Supervision

Motivation A where-to-act mask underdetermines the action: grasping a knife handle, scooping with a spoon, or pouring from a mug each demand a specific end-effector orientation, fixed to first order by the local surface normal at the contact point. Affordance-conditioned VLAs (Sec. 2) [24, 31, 34] delegate this orientation to the downstream policy or a separate grasp planner [4]; we emit it densely, so the grasp packet of Eq. (2) provides a candidate position–orientation cue $g_c = (\mathbf{p}_c, \mathbf{n}_c)$. It is cheap on both counts: the geometry uses the *same* frozen features and forward pass as the mask, and its supervision is free, synthesized from the depth already paired with every affordance dataset.

Ground-Truth Normals from Depth UMD ships no normal annotations; we construct them, following the vertex-map normal estimation of KinectFusion [25] (with a median rather than bilateral pre-filter). Depth is back-projected through the intrinsics, $P(u,v) = Z(u,v)\,K^{-1}(u,v,1)^\top$, and

$$\mathbf{N}^* \propto -\,\partial_u P \times \partial_v P. \tag{15}$$

Finite differencing amplifies Kinect-v1 depth speckle (3–5 mm [15]) into angular noise, so we apply an edge-preserving 5-pixel median filter to Z before differentiation, freezing invalid-depth holes so smoothing cannot hallucinate geometry. Crucially, we also audit the supervision: its raw–filtered consistency discrepancy is 19.2° unfiltered and 15.5° after our filter. This 15.5° quantity is a supervision-consistency estimate, not the model error. Deriving normal supervision from depth this way is established practice in the normal-estimation literature [11,12].

Geometric supervision must remain valid under augmentation. For an in-plane rotation R_ϑ applied to the image, dense targets warp but normal *vectors* must also rotate, and a horizontal flip must negate the lateral component:

$$\mathbf{N}^* \;\to\; \tilde{R}_\vartheta\,(\mathbf{N}^* \circ R_\vartheta^{-1}), \qquad \text{flip: } n_x \to -\,n_x, \tag{16}$$

where $\tilde{R}_\vartheta$ rotates the (n_x, n_y) components. Omitting Eq. (16) yields supervision that is visually plausible but geometrically false; our implementation is verified to $< 0.2^\circ$ round-trip error.

4.4 VLA Probing Protocol

We instantiate Sec. 3 on Ψ_0's stage-1 backbone (Qwen3-VL-2B autoregressively pretrained on egocentric human video; $d = 2048$, $L = 29$; no action expert attached), taking the last-token hidden state as $\mathbf{e}^{(\ell)}$, which dominates mean pooling by over 20 points. We generate 340 goal-level instructions (10 per object–affordance pair, 34 pairs, all 17 UMD categories; none naming the affordance or a synonymous verb), pair each with 4 UMD images of the object (1360 embeddings per condition), and train the linear probe of Eq. (4) on the novel-instruction split (952/408). We compared a 512-unit MLP control with linear access, and report accuracies as mean±std over 3 seeds.

5 Experimental Setup

All segmentation runs use the official data from UMD dataset of Myers *et al.* [23]. Inputs are crops of the source frames, preserving object scale. Full settings are collected in Tab. 1. All the metrics come from the `aff-seg-eval` toolkit of [3] applied to our predictions as defined in Eq. (3). We use the same official category split, test set, and evaluation settings as the baseline results. Its Jaccard accumulates intersections and unions per class over the test set,

$$J_c = \frac{\sum_i |P_c^i \cap G_c^i|}{\sum_i |P_c^i \cup G_c^i|}, \qquad J = \frac{100}{C}\sum_c J_c, \tag{17}$$

Table 1: Training hyperparameters (identical for all runs unless stated).

Optimizer	AdamW, lr 10^{-4}, wd 10^{-4}
LR schedule	2-epoch warmup + cosine decay
Batch / epochs	8 / 25, seeded
Backbone	frozen DINOv2 ViT-B/14 (w/ registers)
Input	448×448 centre crop of 640×480
Augmentation	joint geometric + photometric; normal *vectors* rotated with the image, n_x negated on H-flip
Selection	best val loss on a seeded 30% carve-out of the training split (train 10,376 / val 4,447)

alongside the weighted F-measure F_β^ω [22] ($\beta^2 = 0.3$). Normals are scored using the mean of $\arccos\langle \hat{\mathbf{N}}, \mathbf{N}^* \rangle$ over annotated pixels (ranges: 11.25°/22.5°/30°).

Comparison rows are the fair-protocol scores of [3], re-derived by us from their released per-image outputs. Since those methods differ from ours in code and training details, we additionally train a DeepLabV3-ResNet50 control through our pipeline (same split, crop, loss family; mask-only; fully fine-tuned).

6 Results

6.1 Segmentation Results on UMD

Table 2 reports the main result: 80.29 ± 0.29 Jaccard for the multi-task model and 82.92 for the segmentation-only variant, +4.4 and +7.1 over M2F-AFF; the three seeds converge to near-identical validation loss, mIoU, and normal error (Fig. 2). M2F-AFF, a query-based segmenter fine-tuned end-to-end, still trails by 4.4 despite our model using no detection supervision and no object identity.

The advantage is consistent (Tab. 3, Fig. 3): above the DeepLabV3 control on all seven classes, largest on the imbalanced classes like `scoop` (+17.9) and `contain` (+15.0), where the weighting of Eq. (12) operates, and above M2F-AFF on six of seven. Figure 4 shows held-out predictions carrying the multi-label structure: several affordance channels active on distinct parts of an object.

Table 2: UMD, `aff-seg-eval` toolkit [3], Jaccard ×100. Ours: mean±std (3 seeds).

Method	Jaccard	$F_\beta^\omega \times 100$
CNN [3]	57.44	—
AffordanceNet [10]	63.33	—
DRNAtt [14]	69.50	—
DeepLabV3 [7]	70.19	79.36
M2F-AFF [3] (previous best)	75.86	—
Ours (multi-task, w/ normals)	$\mathbf{80.29 \pm 0.29}$	86.07 ± 0.21
Ours (multi-task, w/o normals, 1 seed)	82.92	**87.79**

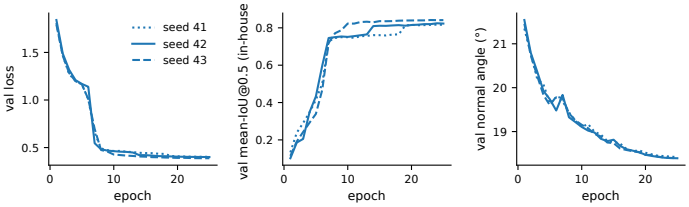

Fig. 2: Metrics validation curves (3 seeds): loss, mean-IoU@0.5, normal angular error.

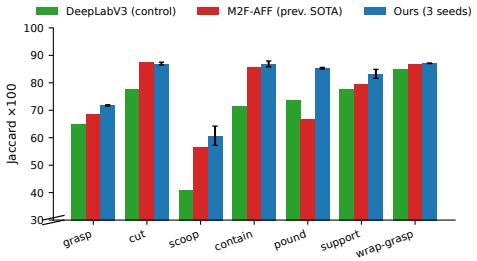

Fig. 3: Per-class Jaccard: ours (3-seed mean±std) vs. M2F-AFF (recomputed from released per-image CSVs) vs. our DeepLabV3 control. Ours ≥ M2F-AFF on 6/7 classes.

Table 3: Per-class Jaccard ×100 (toolkit), test set. The multi-task column is the mean over seeds 41/42/43; the control and segmentation-only columns are seed 42.

Class	DeepLabV3	Ours (w/ normals)	Ours (w/o normals)
`grasp`	64.99	71.77	**72.98**
`cut`	77.69	87.00	**88.44**
`scoop`	40.88	60.72	**70.61**
`contain`	71.51	86.89	**88.47**
`pound`	73.58	85.31	**86.37**
`support`	77.61	83.26	**86.14**
`wrap-grasp`	85.06	87.10	**87.42**
Mean	70.19	80.29	**82.92**

6.2 Surface Normal Estimation

The normals head reaches 21.65° mean angular error (Tab. 4), within a few degrees of the 15.5° depth-supervision consistency estimate measured in Sec. 4.3. It beats the best per-image *constant* normal by ≈ 10°, and the predicted fields are qualitatively smoother than their depth-derived targets. Error is not uniform across the 51 test tools: per-tool mean angular error ranges from 17.1° (`mug`) to 38.3° (`scissors`), a 4.3° standard deviation driven by thin, high-curvature parts (blades, tines) where finite-difference normals are least stable.

Table 4: Surface normal estimation: mean angular error and pixels % within each threshold. The 21.65° mean is reported alongside a 15.5° label consistency estimate.

	Mean err.	≤ 11.25°	≤ 22.5°	≤ 30°
Ours	21.65°	24.1%	62.7%	79.3%

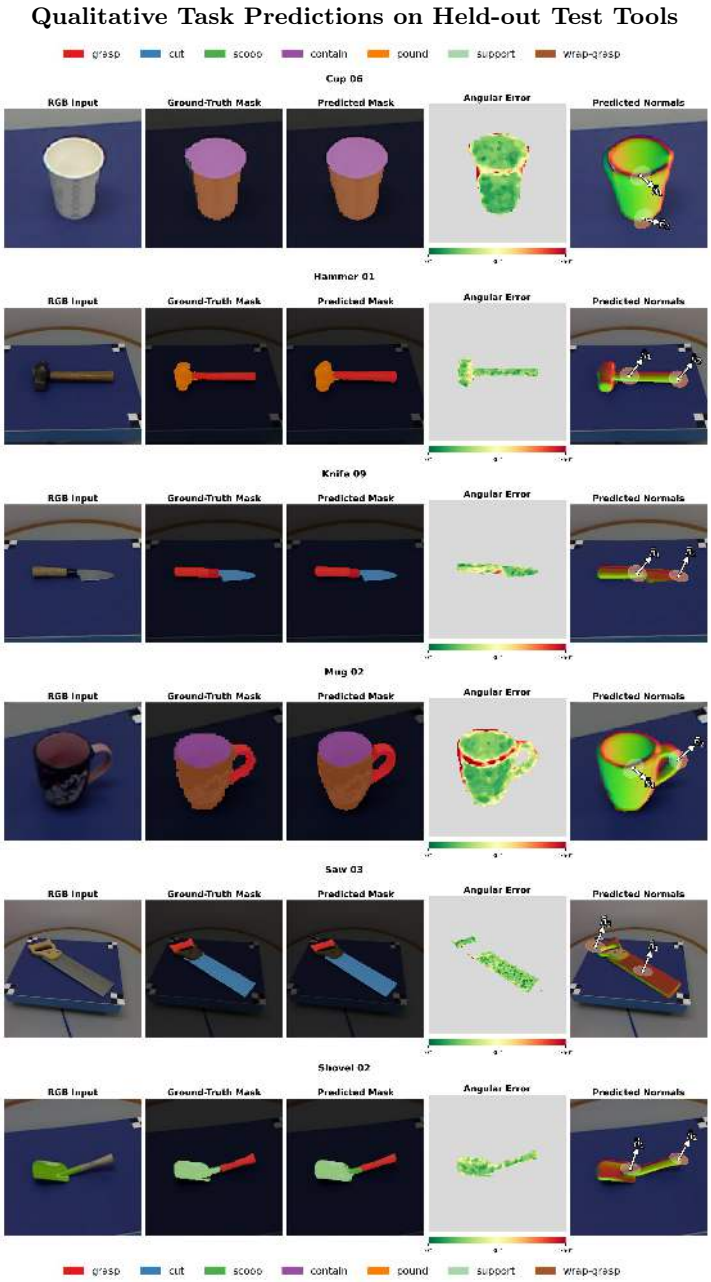

Fig. 4: Held-out test instances, one tool per row: RGB input, ground-truth affordance mask, predicted mask and surface normals (color-coded by direction: the red, green and blue channels give the normal's x, y, z components, *i.e.* left–right, up–down, and depth). Predictions preserve the multi-label structure and stay within object contours, spanning `grasp`, `cut`, `scoop`, `contain`, `pound`, `support`, and `wrap-grasp` affordances.

6.3 Ablation Analysis

Normal Head and Backbone Size At seed 42, dropping the normal head, which corresponds to setting $\lambda_n = \lambda_s = 0$ in Eq. (11), gives 82.92 Jaccard vs. 80.18 with it: the grasp-packet geometry costs 2.74 Jaccard, concentrated in `scoop` (70.6 vs. 58.8). Substituting ViT-S (5.2M params.) already matches M2F-AFF at 76.72 Jaccard, with ViT-B adding +3.5, concentrated in `scoop` and `support` classes.

Object Isolation A natural deployment heuristic isolates the target before prediction. We test it by compositing the object onto neutral gray via a SAM mask seeded by (i) oracle boxes obtained from the ground-truth affordance mask, (ii) Grounding DINO with the object name, (iii) Grounding DINO with a generic "a handheld tool. a utensil. an implement." prompt (Tab. 5).

Table 5: Isolation / localization gradient, test set, Jaccard ×100. The raw-image reference is the 3-seed headline mean; all isolation variants use seed 42.

Localization regime	Jaccard	F_β^ω
None (raw image; headline)	**80.29 ± 0.29**	**86.07 ± 0.21**
SAM, oracle GT affordance box	75.37	82.95
SAM, Grounding DINO, object-name prompt	75.77	83.75
SAM, Grounding DINO, generic prompt	69.55	78.64

We tested it using seed 42 to compare it with the baseline, and the result is that isolation reduces Jaccard under every regime, even oracle boxes (−4.8), though localization quality still matters within the family (generic 69.55 ≪ oracle 75.37). Graying the background removes scene context the frozen features exploit; our headline therefore uses no test-time assistance. A second possibility is that SAM masks out regions that are affordance areas, reducing detection accuracy. The opposite holds on the out-of-distribution kitchenware images, where isolating the object with SAM qualitatively improves predictions on unseen objects (Sec. 7.2).

7 Discussion

7.1 Affordance Grounding in a Humanoid VLA

Layer-wise Decodability At Ψ_0's last layer (Tab. 6) the ordering `full` > `instruction_only` > `image_only` > `chance` holds with $\Delta^{(28)} = +7.4$. The layer sweep resolves three regimes: lexical early layers where the image is a distractor ($\Delta^{(1)} = -8.7$); a mid-depth band where cross-modal integration peaks ($\Delta^{(15..17)} = +9.4$ to $+10.2$): full accuracy is maximal at L15 (84.6 ± 0.7), while the visual gain is maximal at L16 (+10.2). It then partially recedes toward the output, so a potential router built on this signal should therefore tap a mid layer.

Application in VLM Action Pretraining The control column of Tab. 6 is crucial. Base Qwen3-VL scores higher in absolute terms (92.3%), but swept identically across all 29 layers its visual contribution never exceeds +3.3 and is negative in the early layers (−10.4 at L1), whereas Ψ_0's reaches +10.2 mid-stack. Using the same probe, dataset, and split isolates the effect of the backbone: Ψ_0 encodes affordance information in its visual representation, whereas the base VLM does not. Since action pretraining on egocentric video is the only difference between the two models, the observed gap can be attributed to that stage. Linear and MLP probes perform similarly (76.0 vs. 75.9 at L28), indicating that the information is already linearly separable. Rather than suggesting that action pretraining is the only way to acquire this representation, these results point to affordance prediction as a promising auxiliary objective for VLM-to-VLA post-training. Since affordance information is already partially present and becomes more accessible after action pretraining, explicitly supervising it may provide a more data-efficient alternative (student-teacher distillation of affordance models) to relying solely on large-scale egocentric and teleoperation datasets.

Table 6: Last-layer (L28) linear-probe accuracy (%, 3 seeds): action-pretrained Ψ_0 vs. the base Qwen3-VL-2B control. Chance 14.3%. Δ = visual contribution, Eq. (5).

Condition	Qwen3-VL-2B	Ψ_0
full (image+instruction)	92.3 ± 0.8	76.0 ± 1.4
instruction_only	91.8 ± 1.7	68.6 ± 3.2
image_only	40.2 ± 3.1	47.9 ± 0.3
Δ (full − instruction)	+0.5	**+7.4**

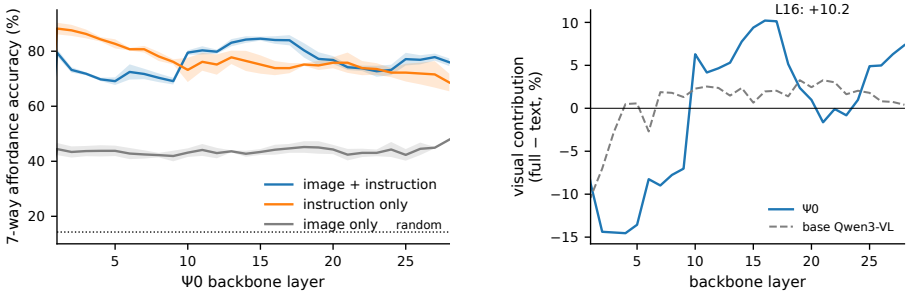

Fig. 5: Layer-wise affordance probing. **Left:** Linear-probe accuracy for Ψ_0 with `image_only`, `instruction_only`, and `full` inputs. **Right:** Visual contribution across layers for Ψ_0 and Qwen3-VL-2B. The stronger mid-layer contribution in Ψ_0 suggests that action pretraining implicitly encodes affordance information.

7.2 Limitations

On 21 real kitchenware images, zero-shot predictions transfer within object contours to object categories absent from UMD (pots, forks, plates). The dominant failure mode is background-texture false positives (*e.g.* `grasp` activating on wood grain), as UMD contains no such negative examples. The pipeline of Sec. 6.3 specifically addresses this isolating the object with Grounding DINO and SAM, which produces qualitatively cleaner predictions than the raw model.

The network predicts on a 448×448 centre *crop* of the 640×480 UMD frame, removing only border pixels while retaining the centred object. The 15.5° raw–filtered discrepancy measures consistency between raw and filtered depth supervision rather than a physical noise floor. Available compute limited the architecture ablations: a preliminary final-layer-only variant was less stable than using four feature taps. A matched-budget DINOv2 fine-tuning comparison and synthetic VLA training on affordances and surface normals remain future work.

8 Conclusion

We presented a two-level study of affordance based on frozen foundation models and lightweight trainable components. A 5.6M-parameter multi-task readout decoder, combining multi-scale token fusion, a from-scratch spatial-detail encoder, and class-frequency-weighted, geometrically equivariant training, sets the state of the art on UMD under the benchmark's official evaluation protocol while predicting local geometry with error reported relative to a depth-supervision consistency estimate. A linear probe provides the task-level counterpart, showing that a frozen humanoid VLA encodes the affordance required by a goal-level instruction, with the visual grounding emerging after action pretraining.

Together, these components provide the information required for downstream manipulation: *(i)* the instruction, *(ii)* the corresponding affordance class, and *(iii)* the associated affordance region and approach normal. This corresponds to the task-level and spatial affordance information that recent policies that use affordances [18, 31, 34] typically learn during policy training, but is recovered here from frozen vision-language models without additional robot data.

Future work will investigate conditioning VLA's action expert on the predicted affordance map, using affordance prediction as an auxiliary objective during VLM-to-VLA post-training, RL fine-tuning with the affordance signal, and evaluation in cluttered RGB-D scenes with closed-loop robot execution.

Acknowledgments

The authors thank Manel Frigola (Universitat Politècnica de Catalunya), as this project originated from work by F. Dorati in his class. It was later extended with T. F. Banfi improving the architecture and including the VLA analysis.

We also thank the Talos Robotics AI (`https://talosrobotics.ai`) research team for the discussions on synthetic VLA training and Stefano Marin (CERN) for his suggestions for improving the visualization of surface-normal predictions.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes (2018), https://arxiv.org/abs/1610.01644
2. Apicella, T., Xompero, A., Cavallaro, A.: Visual affordance prediction: Survey and reproducibility (2025), https://arxiv.org/abs/2505.05074
3. Apicella, T., Xompero, A., Gastaldo, P., Cavallaro, A.: Segmenting Object Affordances: Reproducibility and Sensitivity to Scale, p. 286–304. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-92591-7_18, http://dx.doi.org/10.1007/978-3-031-92591-7_18
4. Atar, S., Li, Y., Grotz, M., Wolf, M., Fox, D., Smith, J.: Optigrasp: Optimized grasp pose detection using rgb images for warehouse picking robots (2024), https://arxiv.org/abs/2409.19494
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631
6. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0: A vision-language-action flow model for general robot control (2026), https://arxiv.org/abs/2410.24164
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017), https://arxiv.org/abs/1706.05587
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1280–1289. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00135, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00135
9. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=2dnO3LLiJ1
10. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 5882–5889 (2018). https://doi.org/10.1109/ICRA.2018.8460902
11. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). p. 2650–2658. ICCV '15, IEEE Computer Society, USA (2015). https://doi.org/10.1109/ICCV.2015.304, https://doi.org/10.1109/ICCV.2015.304
12. Fouhey, D.F., Gupta, A., Hebert, M.: Data-driven 3d primitives for single image understanding. In: 2013 IEEE International Conference on Computer Vision. pp. 3392–3399 (2013). https://doi.org/10.1109/ICCV.2013.421

13. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
14. Gu, Q., Su, J., Yuan, L.: Visual affordance detection using an efficient attention convolutional neural network. Neurocomputing **440**, 36–44 (2021). https://doi.org/https://doi.org/10.1016/j.neucom.2021.01.018, https://www.sciencedirect.com/science/article/pii/S0925231221000278
15. Khoshelham, K., Elberink, S.O.: Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors **12**(2), 1437–1454 (2012). https://doi.org/10.3390/s120201437, https://www.mdpi.com/1424-8220/12/2/1437
16. Kim, T., Bae, H., Li, Z., Li, X., Ponomarenko, I., Wu, R., Dong, H.: Manipgpt: Is affordance segmentation by large vision models enough for articulated object manipulation? (2025), https://arxiv.org/abs/2412.10050
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
18. Kong, W., Su, Z., Yu, W., Dong, H.: Affordvla: Injecting affordance representations into vision-language-action models via implicit feature alignment (2026), https://arxiv.org/abs/2605.17517
19. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding . In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10922–10931. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01051
20. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 38–55. Springer Nature Switzerland, Cham (2025)
21. Ma, T., Zheng, J., Wang, Z., Gao, Z., Zhou, J., Liang, J.: Glover++: Unleashing the potential of affordance learning from human behaviors for robotic manipulation (2025), https://arxiv.org/abs/2505.11865
22. Margolin, R., Zelnik-Manor, L., Tal, A.: How to Evaluate Foreground Maps . In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2014). https://doi.org/10.1109/CVPR.2014.39, https://doi.ieeecomputersociety.org/10.1109/CVPR.2014.39
23. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: 2015 IEEE International Conference on Robotics and Automation (ICRA). pp. 1374–1381 (2015). https://doi.org/10.1109/ICRA.2015.7139369
24. Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., Xiao, T.: Rt-affordance: Affordances are versatile intermediate representations for robot manipulation (2024), https://arxiv.org/abs/2411.02704
25. Newcombe, R.A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A.J., Kohi, P., Shotton, J., Hodges, S., Fitzgibbon, A.: Kinectfusion: Real-time dense surface mapping and tracking. In: 2011 10th IEEE International Symposium on Mixed and Augmented Reality. pp. 127–136 (2011). https://doi.org/10.1109/ISMAR.2011.6092378

26. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Detecting object affordances with convolutional neural networks. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2765–2770 (2016). https://doi.org/10.1109/IROS.2016.7759429
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
28. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12159–12168 (2021). https://doi.org/10.1109/ICCV48922.2021.01196
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
31. Wang, R., Fu, Y., Li, Y., Lin, T., Qian, T., Elhoseiny, M., Zhao, B., Fu, Y., Jiang, Y.G., Xue, X.: Afford-vla: Action-aligned visual planning via internalized affordance (2026), https://arxiv.org/abs/2605.24203
32. Wei, S., Jing, H., Li, B., Zhao, Z., Mao, J., Ni, Z., He, S., Liu, J., Liu, X., Kang, K., Zang, S., Yuan, W., Pavone, M., Huang, D., Wang, Y.: ψ_0: An open foundation model towards universal humanoid loco-manipulation (2026), https://arxiv.org/abs/2603.12263
33. Wu, D., Fu, Y., Huang, S., Liu, Y., Jia, F., Liu, N., Dai, F., Wang, T., Anwer, R.M., Khan, F.S., Shen, J.: Ragnet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping (2025), https://arxiv.org/abs/2507.23734
34. Yu, Q., You, J., Wang, Y., Liang, J., Ping, B., Tian, Y., Chen, Y., Cai, M., Gong, Z., Wu, R., Li, Y., Liang, J., Chen, Y.: Affordancevla: A vision-language-action model empowering action generation through affordance-aware understanding (2026), https://arxiv.org/abs/2606.06155
35. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P.R., Salazar, G., Ryoo, M.S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.W.E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N.J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K.A., Driess, D., Ding, T., Choromanski, K.M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M.G., Han, K.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Tan, J., Toussaint, M., Darvish, K. (eds.) Proceedings of The 7th Conference on Robot Learning. Proceedings of Machine Learning Research, vol. 229, pp. 2165–2183. PMLR (06–09 Nov 2023), https://proceedings.mlr.press/v229/zitkovich23a.html