UAD: Unsupervised Affordance Distillation for Generalization in Robotic Manipulation

Yihe Tang, Wenlong Huang, Yingke Wang, Chengshu Li, Roy Yuan,
Ruohan Zhang, Jiajun Wu, Li Fei-Fei
Stanford University

- Problem / objective
  - 기존 affordance predictions이 manual annotation에 너무 의존적.
- Contribution / Key idea
  - Unsupervised Affordance Distillation (**UAD**)
    - 파운데이션 모델의 affordance knowledge를 task-conditioned affordance model에 distillation하는 방식으로, 자동으로 affordance annotation 획득.

- **Overview**

I.E., (1) UAD distillation learning → (2) Policy imitation learning

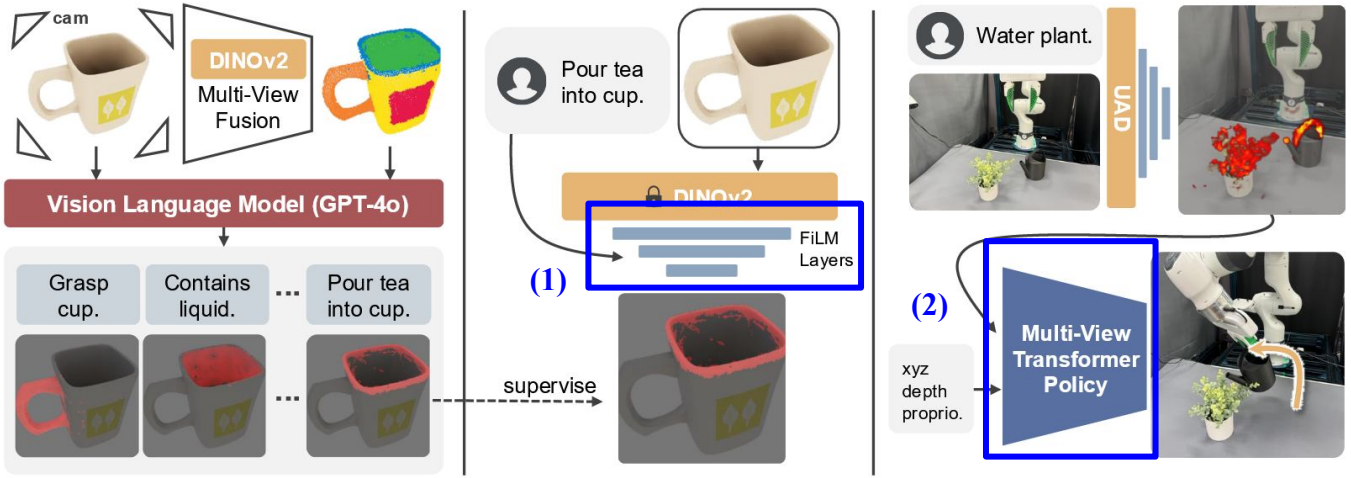

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step1: Extracting Affordance Annotations**

1. 3D objects rendering → 2. Multi-view DINOv2 features fusion → 3. Clustering

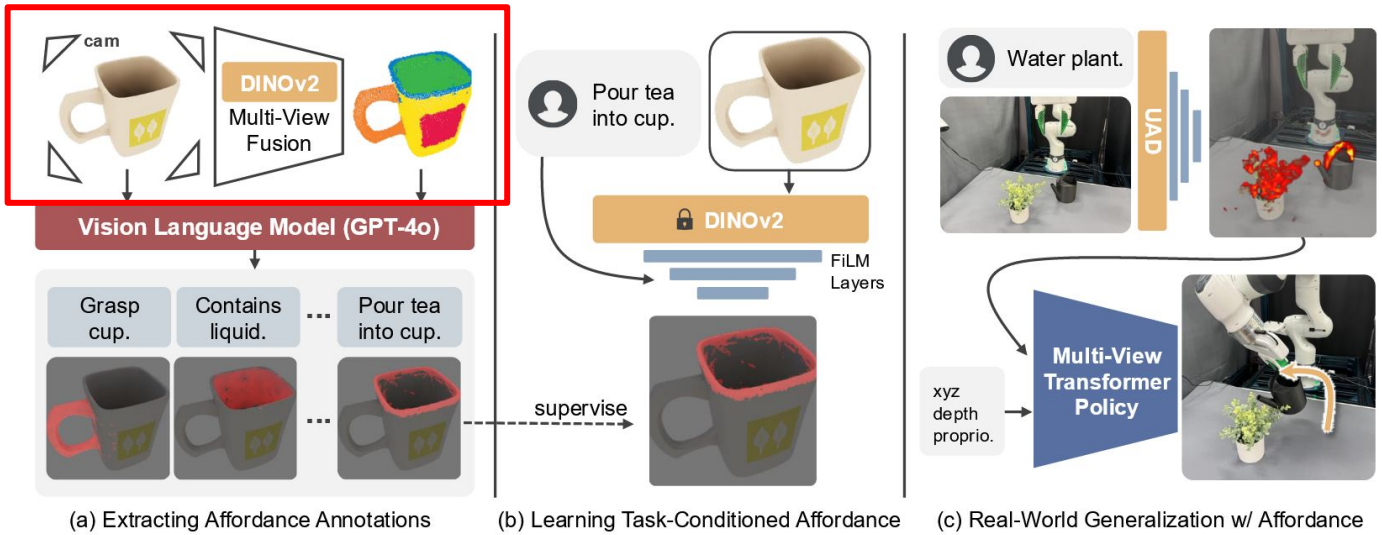

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step1: Extracting Affordance Annotations**

1. 3D objects rendering → 2. Multi-view DINOv2 features fusion → 3. Clustering

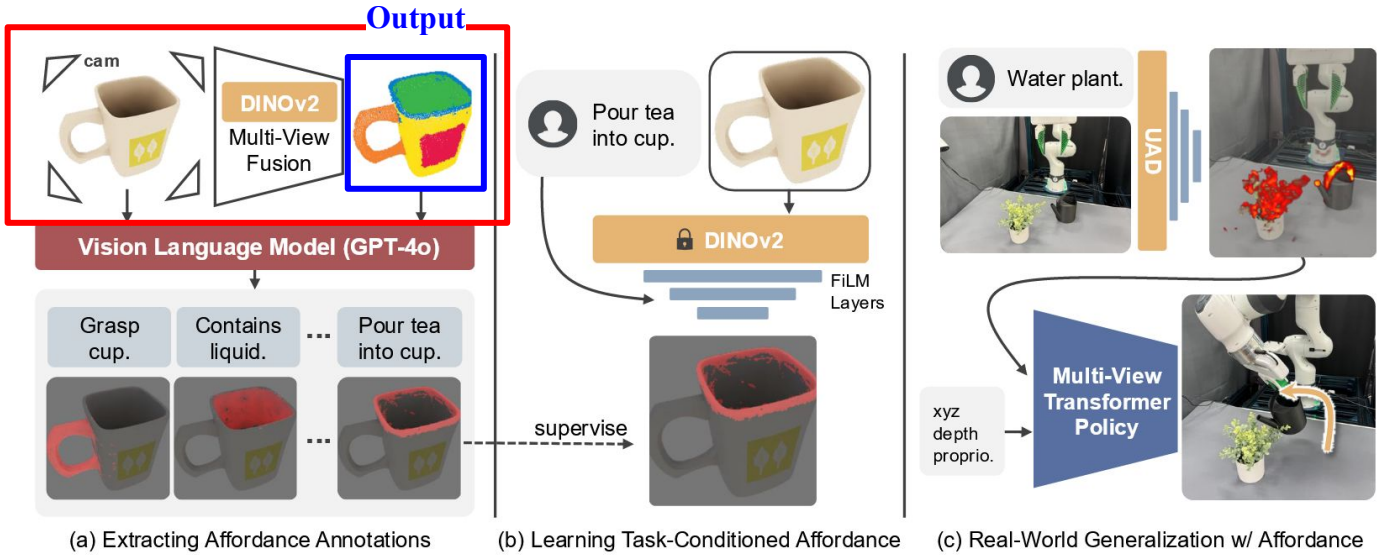

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step1: Extracting Affordance Annotations**

4. Extract affordance annotations via GPT-4o, called $A \in [0,1]^{H \times W}$ as 'GT affordance map'.

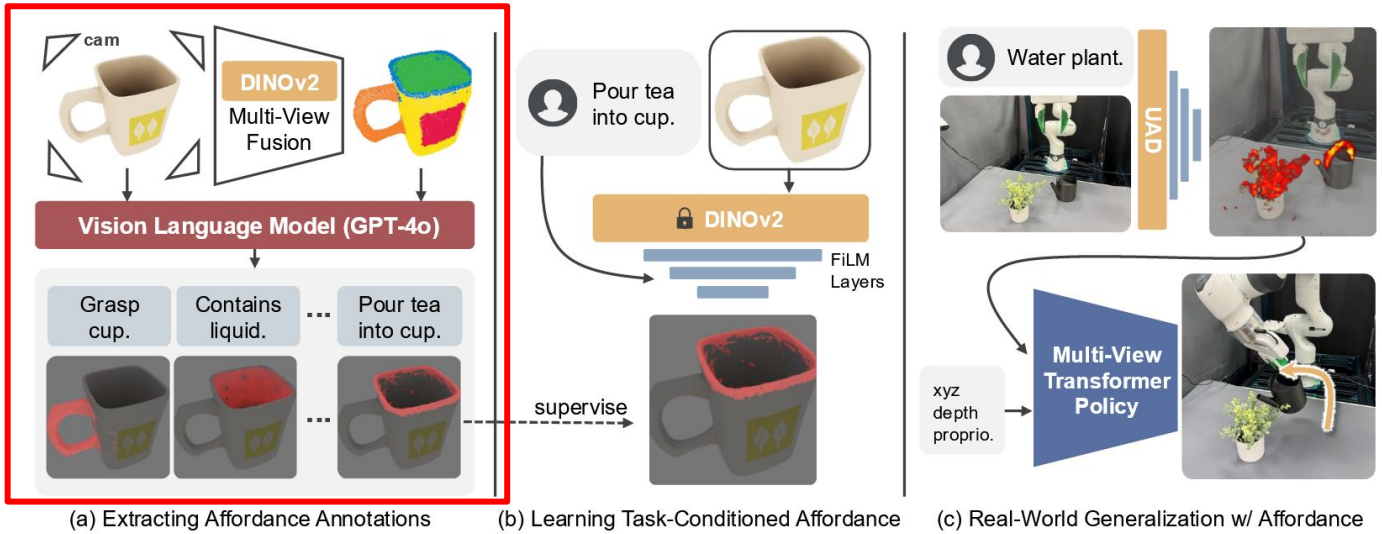

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step1: Extracting Affordance Annotations**

4. Extract affordance annotations via GPT-4o, called $A \in [0,1]^{H \times W}$ as 'GT affordance map'.

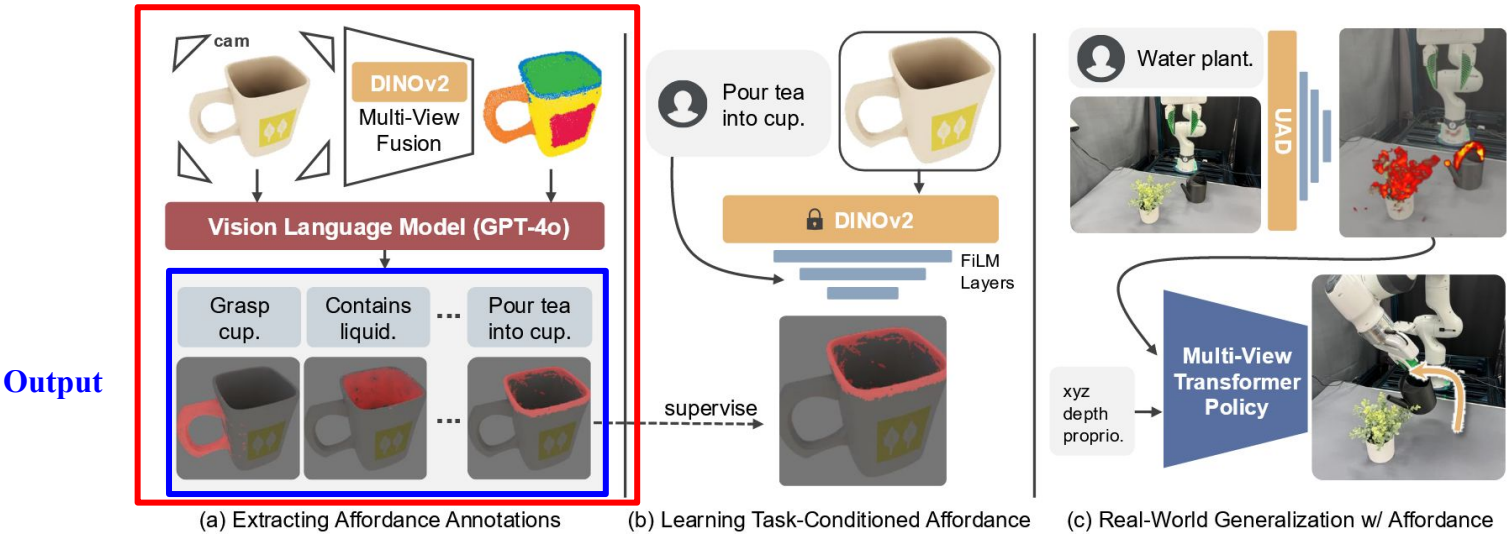

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step2: Learning Task-Conditioned Affordance Model**

1. Train 3 FiLM layers with BCE loss b/w output $\hat{A} \in [0,1]^{H \times W}$ and GT affordance map $A \in [0,1]^{H \times W}$

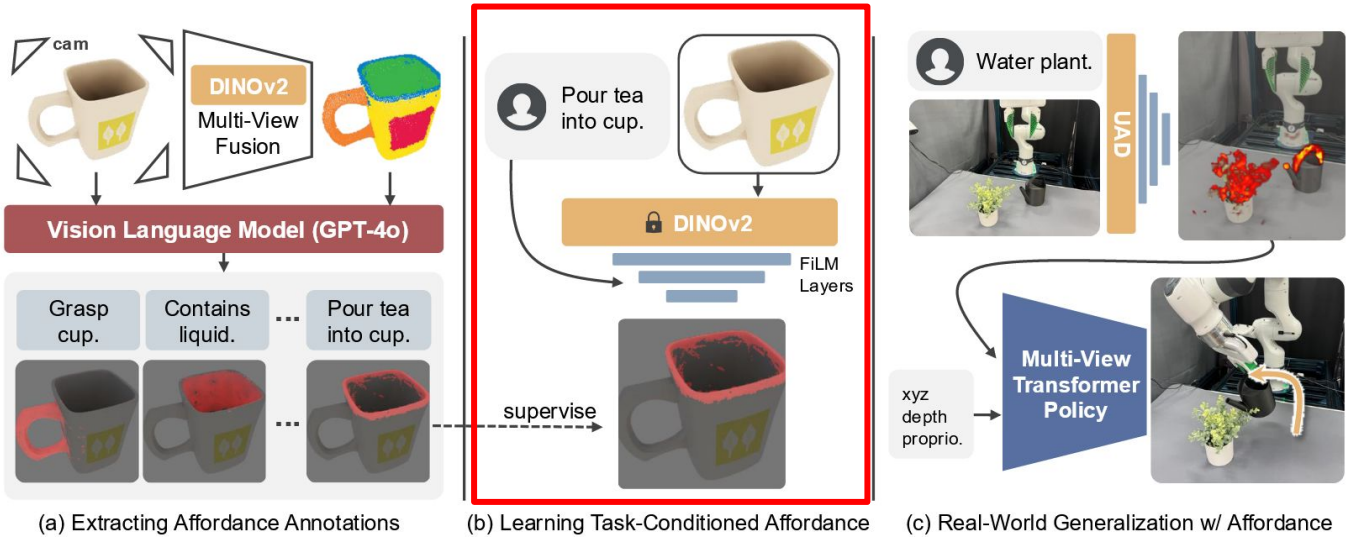

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step2: Learning Task-Conditioned Affordance Model**

1. Train 3 FiLM layers with BCE loss b/w output $\hat{A} \in [0,1]^{H \times W}$ and GT affordance map $A \in [0,1]^{H \times W}$

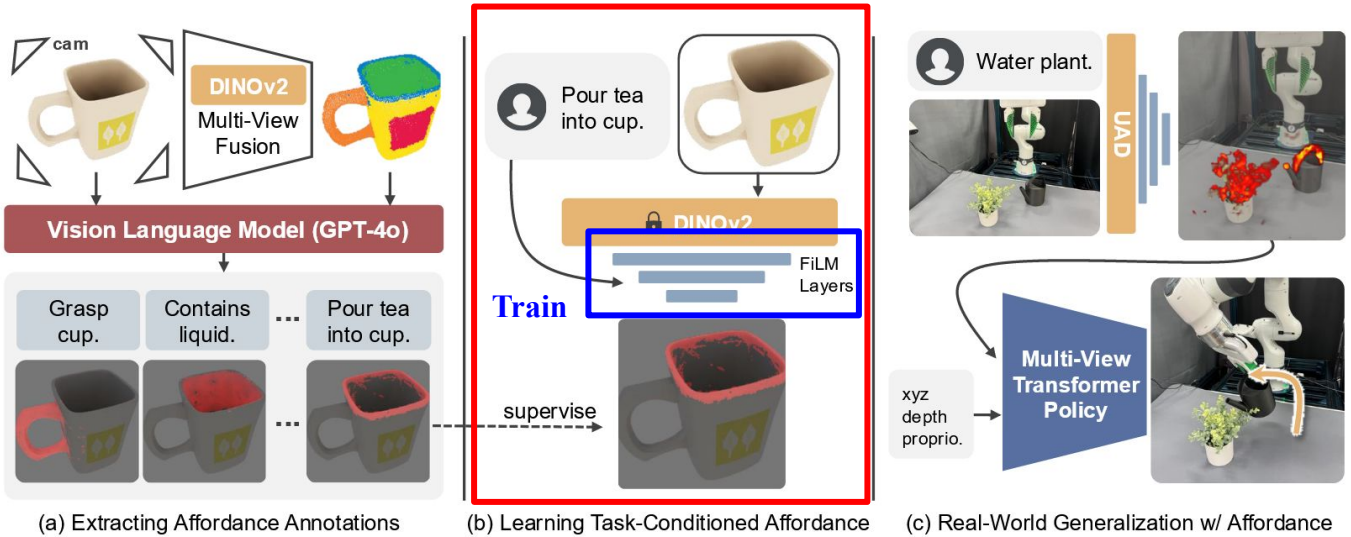

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step3: Policy Learning with Affordance as Observation Space**

1. Predict affordance map

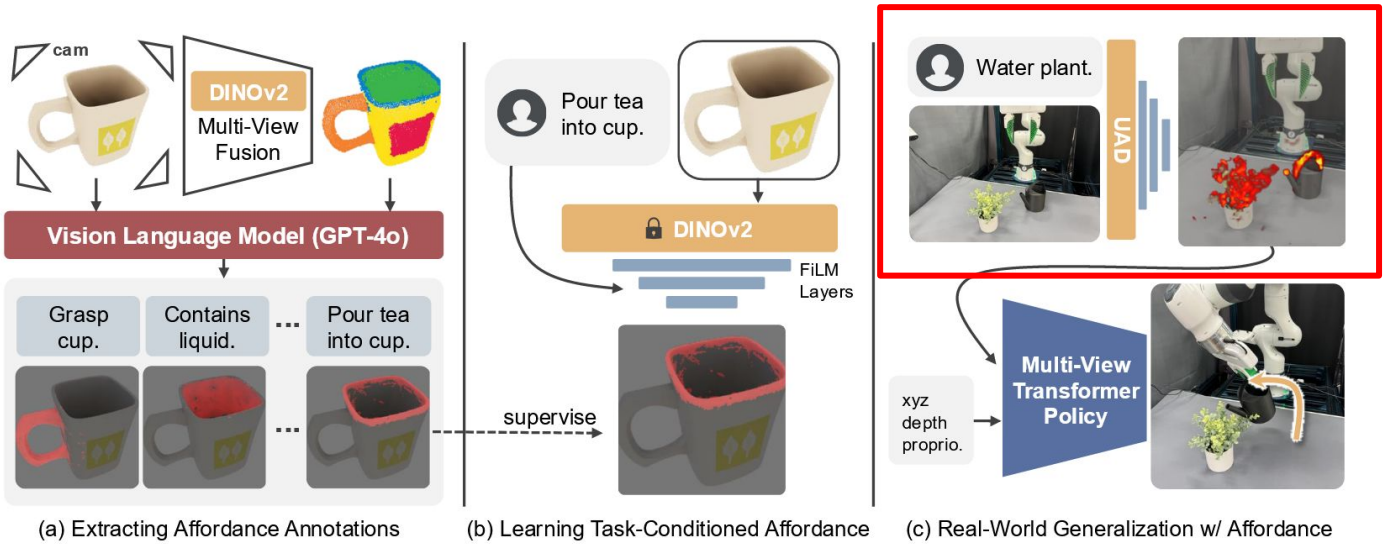

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step3: Policy Learning with Affordance as Observation Space**

1. Predict affordance map

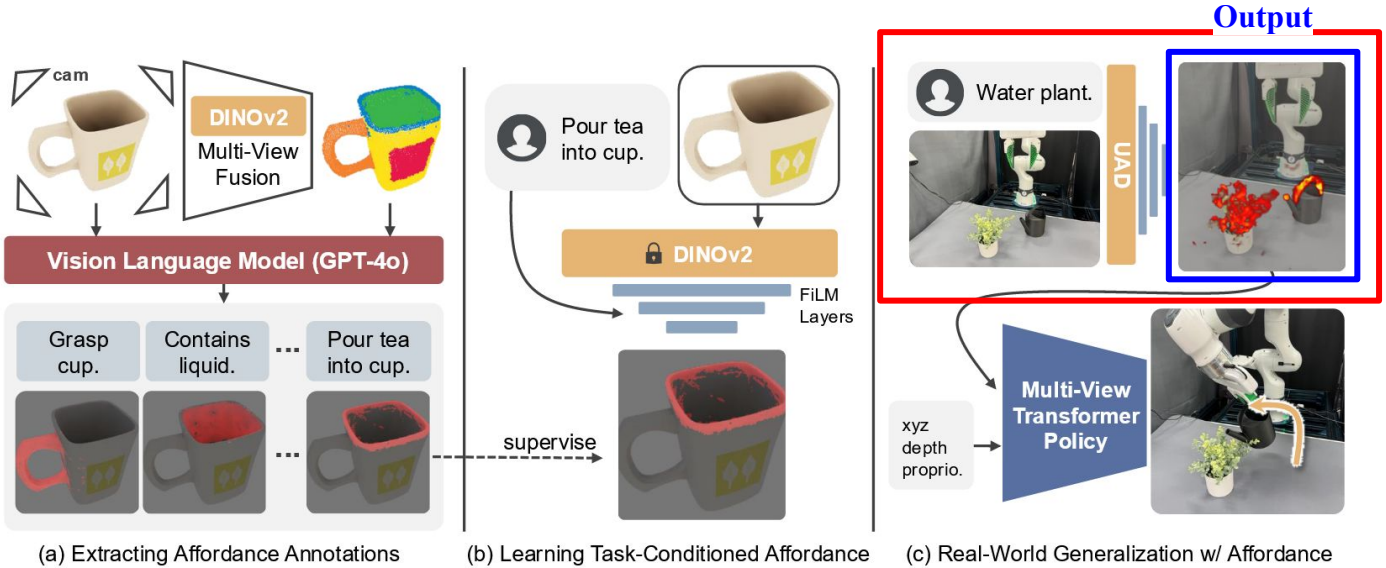

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step3: Policy Learning with Affordance as Observation Space**

2. RVT (Robotic View Transformer) 의 인풋으로 각 view마다 UAD-predicted affordance map도 추가하여 imitation learning.

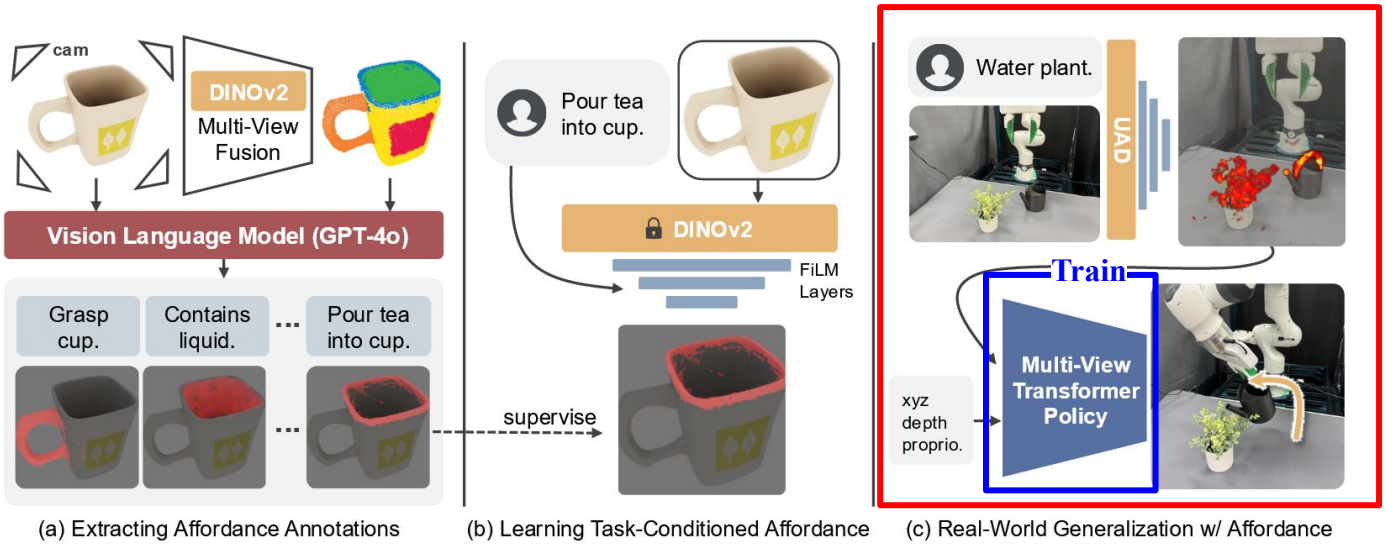

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.

- **Step3: Policy Learning with Affordance as Observation Space**

2. Policy outputs 7-dimension action.

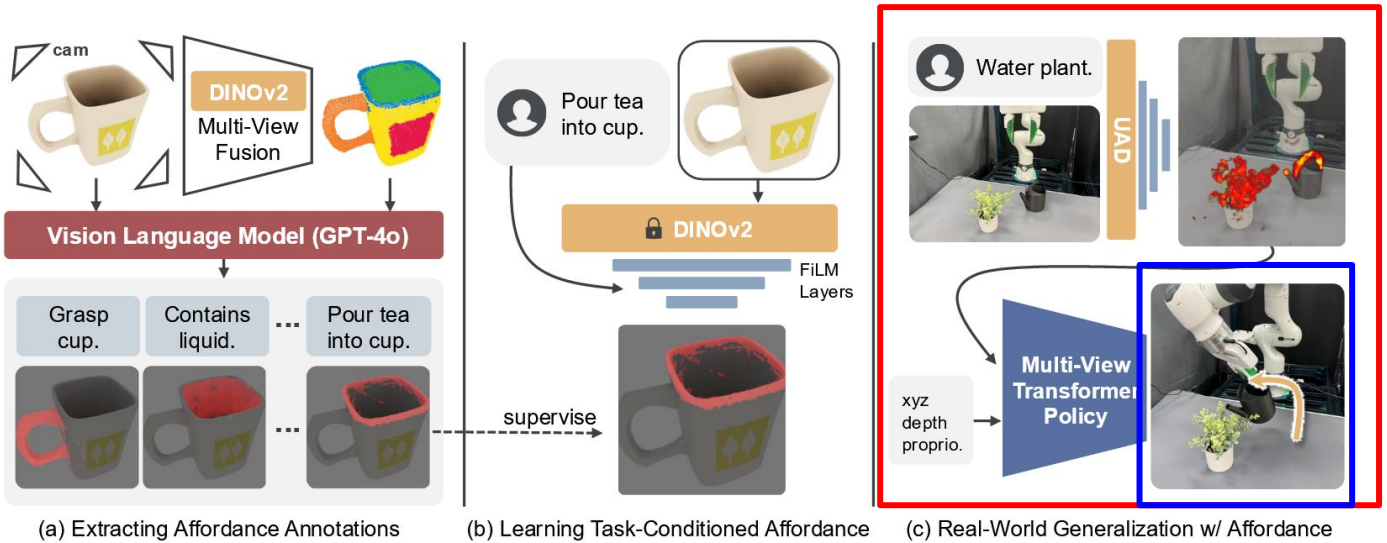

Fig. 2: **Overview of Unsupervised Affordance Distillation (UAD).** Using renderings of 3D objects, we first perform multi-view fusion of DINOv2 features and clustering to obtain fine-grained semantic regions of objects, which are then fed to VLM for proposing relevant tasks and corresponding regions **(a)**. The extracted affordance is then distilled by training a language-conditioning FiLM atop frozen DINOv2 features **(b)**. The learned task-conditioned affordance model provides in-the-wild prediction for diverse fine-grained regions, which are used as observation space for manipulation policies **(c)**.